FUNCTIONAL ANALYSIS OF LOW GRADE GLIOMA GENETIC VARIANTS USING STATISTICS AND PHYSICS-INSPIRED DEEP LEARNING METHODS

BY

JIALU YAN

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Physics
in the Graduate College of the
University of Illinois Urbana-Champaign, 2021

Urbana, Illinois

Doctoral Committee:

Professor Karin A. Dahmen, Chair
Professor Jun S. Song, Director of Research
Associate Professor Sihai Dave Zhao
Assistant Professor Sangjin Kim

Abstract

Large-scale genome-wide association studies (GWAS) have implicated thousands of germline variants in modulating individual's risk of diseases, including cancer. For low grade gliomas (LGGs), at least 25 risk loci have been identified, whose molecular functions, however, remain largely unknown. Understanding how the risk loci function in tumorigenesis poses a major challenge in the field, owing to potential confounding factors and the lack of relevant types of experimental data in the brain. Based on statistical methods and physics-inspired deep learning methods, this work presents a comprehensive computational framework for performing functional analysis of LGG GWAS loci. We hypothesized that GWAS loci contain causal single nucleotide polymorphisms (SNPs) which reside in accessible open chromatin regions and modulate the expression of target genes by perturbing the binding affinity of transcription factors (TFs). We performed an integrative analysis using genomic, epigenomic and transcriptomic data from public repositories and identified the candidate (causal SNP, target gene, TF) triplets that might contribute to oncogenesis. We assessed a candidate causal SNP's potential regulatory role via convolutional neural network (CNN) and simulated-annealing-based interpretation methods. Finally, we applied tensor train decomposition (TT-decomposition) to neural network parameter reduction and demonstrated that the reduced convolutional neural network performed well. This work helps understand the molecular mechanisms underlying genetic risk factors of low grade glioma. The CNN and TT-decomposition-based deep learning approach may benefit future functional genomic studies, where TF chromatin immunoprecipitation followed by sequencing (ChIP-seq) data are not readily available in the brain.

To my parents.

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my advisor, Professor Jun S. Song. Professor Song led me into the genomics and deep learning fields, and helped push the best potential out of me. During the past years in Professor Song's group, I was not only trained on how to do research rigorously, but also on critical thinking, presentation and academic writing skills, all of which helped shape me into a better scientist. In Professor Song's group, I also got exposed to advanced math ideas and computational methods, which greatly broadened my knowledge breadth. I sincerely thank Professor Song for his guidance, patience and encouragement during my graduate study period. I also thank Jun for his advice and continuous support during my internship and on pursuing my career path after the graduate school. The high standard requirement, hard-working attitude, open-mindedness to new ideas and fearlessness to difficult problems will continuously benefit me in the future.

Second, I would like to thank my committee members, Professor Karin A. Dahmen (Chair), Professor Sihai Dave Zhao and Professor Sangjin Kim, for their kindly guidance, suggestions and encouragement for my research. The questions raised by my committee members during my preliminary exam provided valuable advice for the project, and most importantly, helped expose my weakness. Professor Dahmen let me realize the importance of academic writing, and provided precious advice and resources for improving my academic writing skills, which I really appreciated.

Next, I would like to thank Dr. Mohith Manjunath and Dr. Yi Zhang, who gave me tremendous help when I started working on the low grade glioma project. I also want to thank all my collaborators: Professor Robert B. Jenkins (Mayo Clinic) for his valuable advice during our collaboration meetings; Professor Paul R. Selvin (UIUC), Yeoan Youn (UIUC), Dr. Kristen L. Drucker (Mayo Clinic), Professor Valter Zazubovich (Concordia University) for the experimental validation; Professor Joseph F. Costello (UCSF) and Andrew M. McKinney (UCSF) for helpful discussion and suggestions.

I would like to thank the funding agencies, an Oligodendroglioma Research Award from the National Brain Tumor Society (NBTS), a generous gift from the Dabbiere family, and NIH [R01CA163336] for supporting my research.

I would like to thank all my teachers and mentors who inspired and helped me along my way to this point, in particular, Xiaoyan Hou, my physics teacher in middle school; Yuping Wang, my physics teacher in high school; Professor Liqing Chen, Professor Qingyuan Jin and Dr. Zhifan Zhou in East China Normal University; Professor Ping Yu, Professor Carsten A. Ullrich, Professor Paul F. Miceli, Professor Aigen Li and Professor Ioan Kosztin in University of Missouri-Columbia. I would like to thank Professor S. Lance Cooper, who has always been caring, thoughtful and supportive; Wendy Renee Wimmer, who gave me a lot of help during my Ph.D study period.

I would like to thank all the previous and current members in Professor Song's lab for creating the collaborative, supportive and friendly working environment. They are: Yidong Chen, Alex Finnegan, Miroslav Hejna, Hu Jin, Minji Kim, Somang Kim, Yeonsung Kim, Jacob R. Leistico, Alan Luu, Mohith Manjunath, Timothy Miller, Wooyoung Moon, Steve Yeo, Jimmy Yuan, Shuyi Zhang, Yi Zhang and Chenchao Zhao. In particular, previous group members such as Hu, Alex and Yi set examples for what excellent graduate students should be like, which motivated me a lot during my past years in the lab. I also would like to thank my group colleague as well as good friend Shuyi, for encouraging me during my down moments, and for all the enjoyable and inspiring discussions we had.

I would like to thank all my friends who have always been by my side during the years in China and in the US. Your continuous help and support filled my days with pleasure and passion, and inspired me to be a better person. In particular, I would like to thank my close friend Shengxuan Xu, who has always been there for me when I need the most, and has always been understanding, supportive and thoughtful since we met in the first year of middle school.

Last but not least, I would like to thank my parents, Chenxi Yang and Feng Yan, for their unconditional love and support. They always provide me with the best, and back me up in every decision I have made. They are the source of strength and comfort when I come across difficulties, and I could not make to this point without their devotion and support. I am grateful to my parents for everything they have done for me, and this thesis is dedicated to them.

Contents

List of Abbreviations

ASE	Allele Specific Expression
ATAC-seq	Assay for Transposase-Accessible Chromatin using sequencing
CNN	Convolutional Neural Network
ChIA-PET	Chromatin Interaction Analysis by Paired-End Tag Sequencing
ChIP-qPCR	Chromatin Immunoprecipitation coupled with quantitative PCR
ChIP-seq	Chromatin Immunoprecipitation followed by sequencing
DNase-seq	DNase I hypersensitive sites sequencing
EMSA	Electrophoretic Mobility Shift Assay
ENCODE	Encyclopedia of DNA Elements
EUR	European
eQTL	expression Quantitative Trait Loci
FDR	False Discovery Rate
GBM	Glioblastoma Multiforme
GEO	Gene Expression Omnibus
GWAS	Genome-Wide Association Study
HOSVD	Higher-Order Singular Value Decomposition
H3K4me1	Histone H3 lysine 4 mono-methylation
H3K4me3	Histone H3 lysine 4 tri-methylation
H3K27ac	Histone H3 lysine 27 acetylation
LD	Linkage Disequilibrium
LGG	Low Grade Glioma
MAF	Minor Allele Frequency
MAPQ	MAPping Quality
mRNA	messenger RNA
mtDNA	mitochondrial DNA

NGS	Next-Generation Sequencing
NIH	National Institutes of Health
OR	Odds Ratio
PLAC-seq	Proximity Ligation-Assisted ChIP-seq
PSSM	Position-Specific Scoring Matrix
PWM	Position Weight Matrix
REMC	Roadmap Epigenomics Mapping Consortium
RNA-Seq	RNA sequencing
RNAi	RNA interference
RSEM	RNA-Seq by Expectation Maximization
SNP	Single Nucleotide Polymorphism
SVD	Singular Value Decomposition
shRNA	short hairpin RNA
TAD	Topologically Associating Domain
TCGA	The Cancer Genome Atlas
TF	Transcription Factor
TSS	Transcription Start Site
TT-decomposition	Tensor Train decomposition
TT-format	Tensor Train format
WHO	World Health Organization

Chapter 1

Introduction

1.1 Low grade glioma and the functional genomics

1.1.1 Human low grade glioma (LGG) and the *IDH* mutation

Glial cells are non-neuronal, supportive cells in the central nervous system and the peripheral nervous system [1]. Astrocytes, oligodendrocytes, ependymal cells and microglia are the glial cells in the central nervous system; Schwann cells and satellite cells are the glial cells in the peripheral nervous system. Gliomas are tumors originating in the glial cells of the brain. About 30% of all brain and central nervous system tumors are gliomas [2]. According to the 2016 World Health Organization (WHO) classification of tumors of the central nervous system, gliomas include grade II diffuse astrocytomas, oligodendrogliomas, oligoastrocytomas; grade III anaplastic astrocytomas, anaplastic oligodendrogliomas, anaplastic oligoastrocytomas; and grade IV glioblastomas [3]. Glioblastoma is also called glioblastoma multiforme (GBM), and is an aggressive type of cancer. Low grade glioma (LGG) mainly includes diffuse astrocytic and oligodendroglial tumors [3], and is less aggressive compared to GBM. Patients with low grade glioma have overall better survival. The 2016 WHO classification further incorporated molecular features such as the mutations in either isocitrate dehydrogenase 1 (*IDH1*) or isocitrate dehydrogenase 2 (*IDH2*) (collectively referred to as IDH^{mut}) and codeletion of the chromosome arms 1p and 19q (1p/19q codeletion). By including the status of telomerase reverse transcriptase (*TERT*) promoter mutations, gliomas can be further classified into five main molecular groups based on the presence or absence of the three molecular alterations [4]. The five molecular groups are "*TERT* promoter mutation only", "IDH^{mut} only", "*TERT* promoter and IDH^{mut}", triple-positive (IDH^{mut}, *TERT* promoter mutant, 1p/19q codeleted) and triple-negative (*IDH* wild-type, *TERT* wild-type, 1p/19q non-codeleted). The triple-positive and "IDH^{mut} only" groups compose the majority of LGGs, while "*TERT* promoter mutation only" is prevalent in glioblastoma multiforme [4]. This work considers LGGs only, excluding GBM, with a focus on the triple-positive and "IDH^{mut} only" groups, which are usually oligodendrogliomas and astrocytomas, respectively, in terms of the 2016 WHO classification.

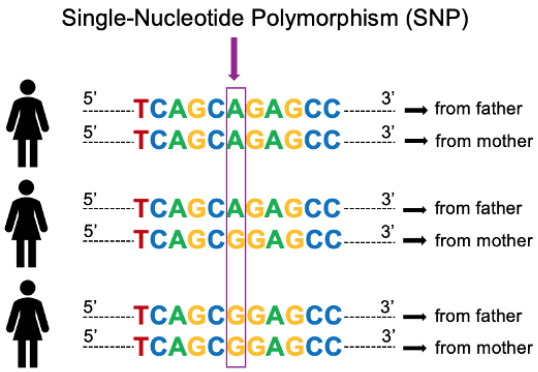
Isocitrate dehydrogenases (*IDHs*) are important enzymes which catalyze the oxidative decarboxylation of isocitrate to α-ketoglutarate (α-KG) [5]. *IDH1* and *IDH2* mutations have been observed in several types of cancer, including glioma and human acute myeloid leukemia. In gliomas, *IDH1* mutations typically involve the amino acid substitution at codon 132, and are the most important and frequent mutations in glioma. Furthermore, *IDH1* mutation is considered as a driver mutation and is usually the first hit in the development of astrocytomas and oligodendrogliomas [6]. The mutant IDH enzymes function by producing 2-hydroxyglutarate (2-HG) from α-KG aberrantly, leading to a global hypermethylation phenotype [7, 5, 8]. One study has shown that 19% of the analyzed 365,092 CpG sites were hypermethylated in IDH^{mut} gliomas compared to *IDH* wide-type gliomas [8]. Hypermethylation is known to alter the global transcription; furthermore, it was also indicated to impact the chromatin topology. W. A Flavahan *et al.* showed hypermethylation at CCCTC-binding factor (CTCF) binding sites in IDH^{mut} gliomas, which led to reduced CTCF binding [9]. CTCF is known to function as an insulator - it creates boundaries of domains across the genome; elements inside those domains are considered to interact more frequently. Reduced CTCF binding could alter the chromatin topology, and leads to aberrant gene expression. Specifically, the study showed due to the loss of CTCF binding at a domain boundary, a constitutive enhancer interacted with an important glioma oncogene platelet derived growth factor receptor alpha (PDGFRA), leading to its increased expression [9]. All above have shown the huge impact caused by *IDH* mutations in gliomas. Given the importance and potential impact of *IDH* mutations in low grade gliomas, we thus focus on the triple-positive and "IDH^{mut} only" subgroups in this study.

1.1.2 Genome-wide association study (GWAS)

Genome-wide association study (GWAS) is an observational study of genetic variants in large populations to find out if any variant is associated with a trait. The genetic variant in a GWAS is usually a single-nucleotide polymorphism (SNP), defined as the single nucleotide germline variant that occurs naturally at some location of the genome (Figure 1.1). The traits in a GWAS are usually human diseases, including cancer.

The first step of the genome-wide association study is to set up the case and control groups, and obtain the allele counts for each SNP in the two groups. Then, to investigate if the allele frequency of one SNP is significantly altered between the case and the control groups [10], odds ratio (OR) is utilized. The odds ratio is defined as the ratio of the odds of one allele and the odds of another allele. Taking Figure 1.1 as an example, the count of allele A in cancer (case) and healthy (control) groups are 30000 and 40000 respectively, and the count of allele G in cancer (case) and healthy (control) groups are 40000 and 30000 respectively.

The odds ratio for allele G in this example is (40000/30000)/(30000/40000). From the definition of odds ratio, we could see that if one allele of a SNP is enriched in the case group compared to the control group, the odds ratio for this allele is larger than 1; conversely, if the frequency of one allele is higher in the control group compared to the case group, the odds ratio for this allele is smaller than 1.



2 x 2 contingency table

	Cancer	Healthy
A (count)	30000	40000
G (count)	40000	30000

Figure 1.1: An illustrative plot of single-nucleotide polymorphism (SNP) and an example of a 2 × 2 contingency table in the genome-wide association study. The left panel: the base in the purple box represents a SNP with Adenine (A) and Guanine (G) as its two alleles. For allele A of the SNP, it can be substituted by allele G in either one chromosome or both chromosomes, resulting in AA, AG and GG genotypes. The right panel: a 2×2 contingency table showing the allele count of A and G in cancer (case) and healthy (control) groups.

Furthermore, people use P-value from the chi-squared test to assess the significance of the odds ratio of a single SNP, and plot $-\log_{10}(P\text{-value})$ of every investigated SNP in a Manhattan plot (Figure 1.2). The Manhattan plot depicts the SNPs that are significantly associated with a trait. The convention of the P-value threshold to call a SNP significant is 5×10^{-8}, due to hundreds of thousands of SNPs tested [10, 11].

To be noted, the significant SNPs discovered by the GWA studies are typically of low penetrance, and are associated with a small increased risk of diseases. In our study, the median odds ratio for the 25 LGG GWAS SNPs was 1.2, where 23 of the 25 SNPs had odds ratio less than 1.5, typical of low-penetrance genetic variants [12] (Appendix A.1).

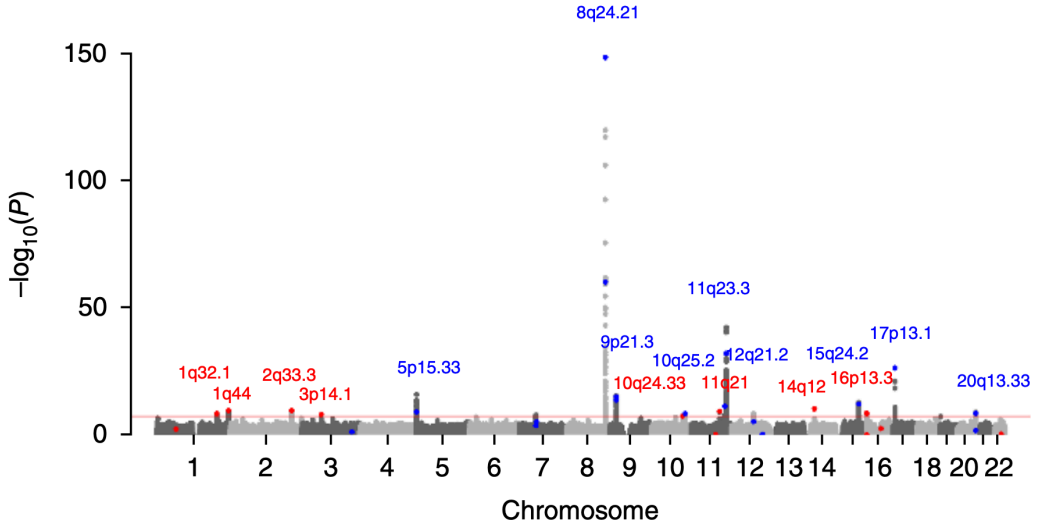


Figure 1.2: An illustration of the Manhattan plot showing several loci significantly associated with LGG. Figure is from Melin *et al.* [12]. The Y-axis is $-\log_{10}(P\text{-value})$ from the genome-wide discovery-phase meta-analysis [12], and the X-axis is the genomic location. Each dot in the figure represents a SNP; larger $-\log_{10}(P\text{-value})$ indicates stronger association. The red horizontal line denotes the P-value threshold 5×10^{-8}. The LGG GWAS loci are highlighted by blue text and red text in the figure, where the red text denotes the newly identified loci from Melin *et al.* [12].

1.1.3 The functional non-coding genome

The "central dogma" in molecular biology describes the flow of genetic information from DNA to RNA, then to protein. The process of copying a segment of DNA to RNA is called transcription, and the process of synthesizing amino acid sequences from RNA is called translation. In human, the whole genome has 23 pairs of chromosomes and around 3 billion pairs of nucleotides. Among the 3 billion DNA base pairs, only 1% - 2% are protein coding regions, while the rest of the genome is the non-coding part. The non-coding genome contains DNA sequences encoding certain types of RNAs, such as microRNAs (miRNAs) and long non-coding RNAs (lncRNAs), all of which have important functions in the cell biological process. Moreover, the non-coding genome contains various types of regulatory elements, for example, promoters, enhancers, silencers and insulators, all of which play an essential role in regulating transcription. Promoters are typically located around the transcription start site (TSS) of genes, and provide binding sites for transcription factors and RNA polymerase to initiate transcription from DNA to RNA (Figure 1.3). Enhancers can be located upstream or downstream of the regulated genes, sometimes even far away from their regulated genes. Enhancers also provide binding sites for transcription factors, which modulate gene expression through interacting with the protein complexes at the promoters (Figure 1.3). Promoters and enhancers are the most well studied

cis-regulatory elements (CREs) [13, 14].

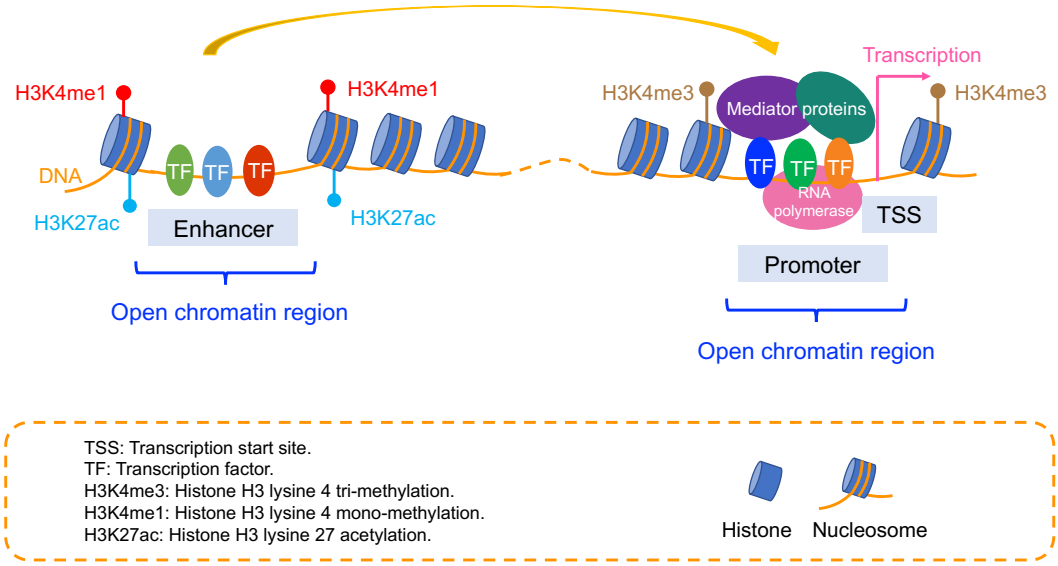


Figure 1.3: An illustrative plot of two functional elements in the genome: promoter and enhancer. Promoters and enhancers reside in open chromatin regions. Promoters could be marked by histone modification H3K4me3; enhancers are usually marked by histone modifications H3K4me1 and H3K27ac. Enhancers can interact with the protein complexes at the promoters to regulate gene transcription. Figure created based on Calo and Wysocka [15].

Promoters and enhancers are marked by epigenetic modifications of the DNA packing proteins - histones. Promoters are marked by the tri-methylation at the 4^{th} lysine residue of the histone H3 protein (H3K4me3), while enhancers are usually marked by the mono-methylation at the 4^{th} lysine residue of the histone H3 protein (H3K4me1) and the acetylation of the 27^{th} lysine residue of the histone H3 protein (H3K27ac) (Figure 1.3). The genomic regions with histone modifications could be determined by one of the high-throughput sequencing methods, named chromatin immunoprecipitation followed by sequencing (ChIP-seq). ChIP-seq is also used to identify the binding sites of transcription factors (Section 1.1.4). In the ChIP-seq procedure, the complexes of DNA segments and their bound transcription factors (TFs), or the complexes of DNA and the chemically modified histone proteins (e.g., H3K4me1) are selectively pulled down by antibodies. The DNA segments are then subjected to high-throughput sequencing to identify the transcription factor binding sites or the sites with certain histone modification [16, 17] (Figure 1.4).

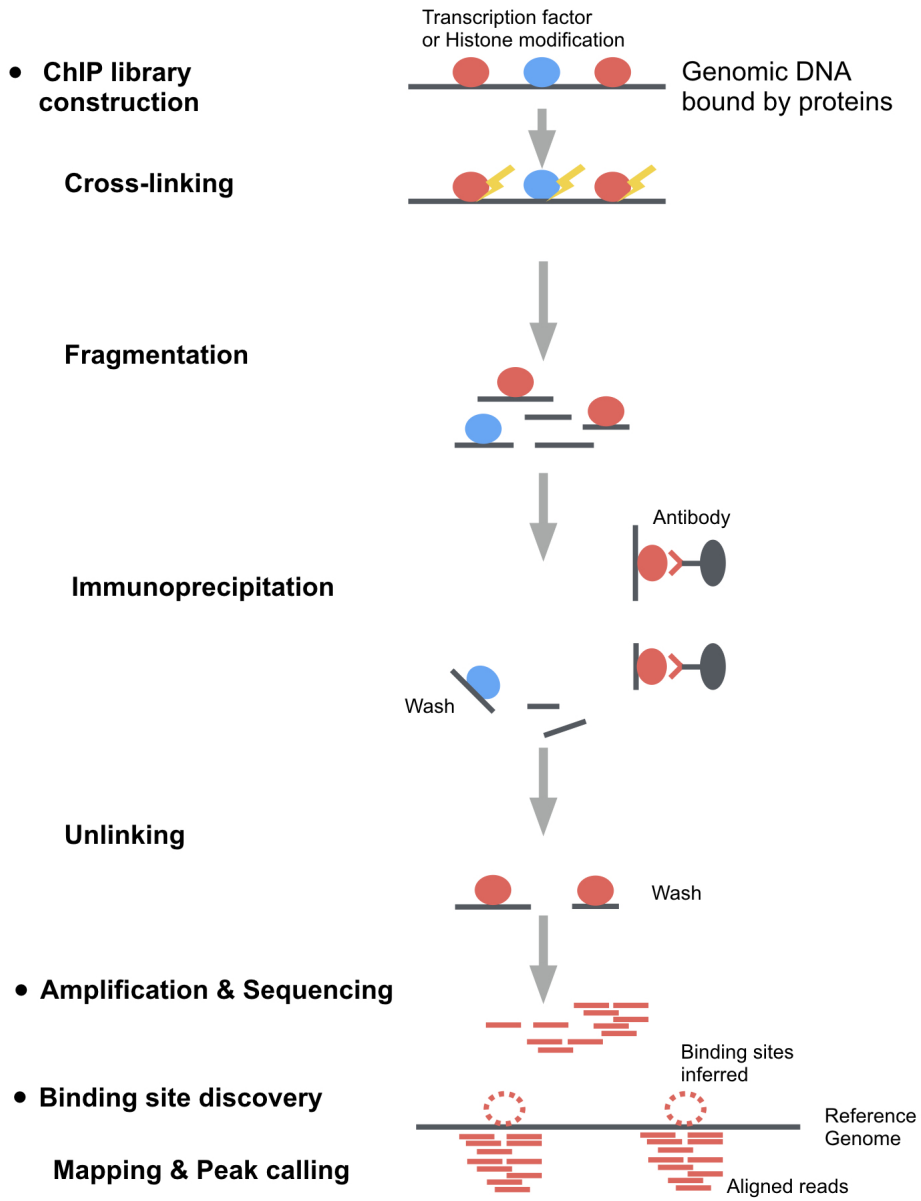


Figure 1.4: An illustrative plot of the ChIP-seq procedure to identify the genomic locations of histone modification or transcription factor binding sites. Figure from SequencEnG [18].

The promoters and enhancers reside in open chromatin regions (Figure 1.3), which can be identified by methods such as DNase I hypersensitive sites sequencing (DNase-seq) and assay for transposase-accessible chromatin using sequencing (ATAC-seq). In DNase-seq, we use DNase I enzyme, which preferentially cleaves open chromatin regions, to obtain the DNase I digested DNA fragments, and then sequence them through

high-throughput sequencing [19] (Figure 1.5A). For ATAC-seq, the open chromatin regions are obtained through inserting the adapters using Tn5 transposase [20] (Figure 1.5B).

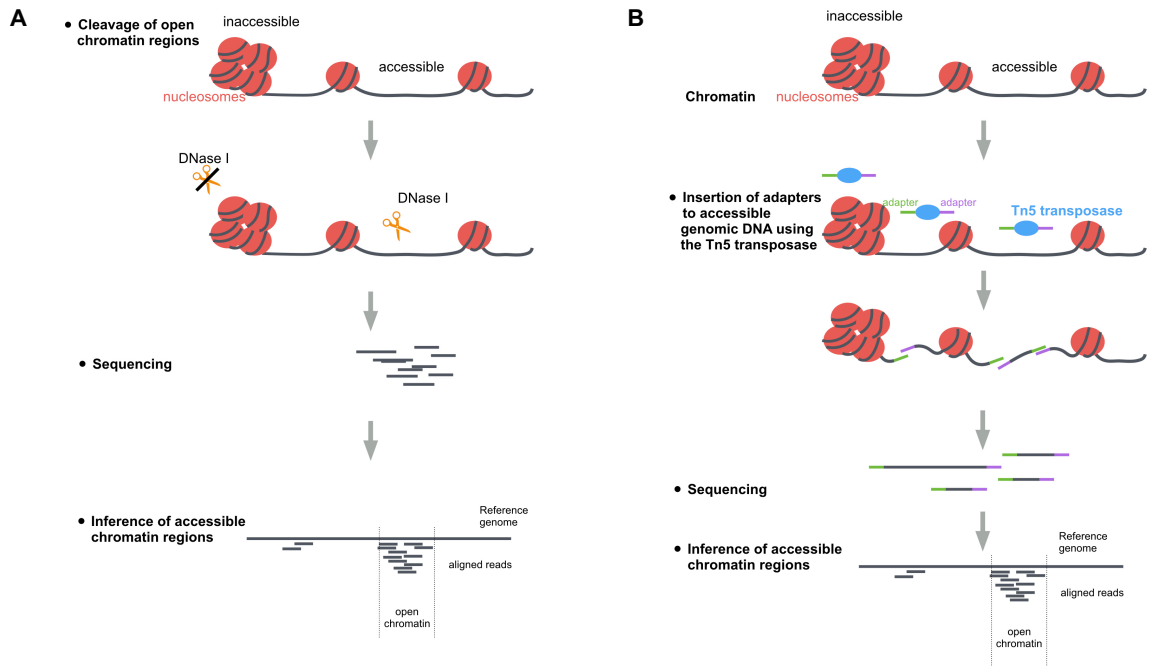


Figure 1.5: An illustrative plot of (A) the DNase-seq and (B) the ATAC-seq methods to identify the open chromatin regions in the genome. Figures are from SequencEnG [18].

As we know, the human chromatin folds to form 3D structures, which indicates that enhancers located hundreds of thousands bases away may be spatially close to the promoters, and thus interact with them. In order to capture the 3D structure of chromatin or the looping of chromatin mediated by a protein, Hi-C [21], proximity ligation-assisted ChIP-seq (PLAC-seq) [22] and chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) [23] could be used.

It is crucial to understand that different cell types may have very divergent landscapes of the functional non-coding genome. For example, the open chromatin regions and the transcription factor binding sites of a glial cell could be very different from those of a blood cell. The functional analysis of the cell-type specific non-coding regions has thus drawn tremendous research interest. Our work in this thesis is also based on the functional annotations of the non-coding genome of brain cells.

1.1.4 Transcription factor (TF) and the DNA binding motif

Transcription factors (TFs) are proteins that bind to DNA sequences and regulate the transcription of genes. TFs include activators and repressors, which function by promoting or blocking the transcription,

respectively. The binding sites of transcription factors in the genome can be experimentally determined by ChIP-seq. TF binding is highly sequence selective, i.e., different TFs recognize different DNA sequences (referred to as motifs). The TF binding motif is often represented by a position-specific scoring matrix (PSSM). The elements in the PSSM matrix are calculated as log likelihoods. Suppose a transcription factor recognizes a motif of length L, the PSSM matrix is:

$$PSSM = \begin{bmatrix} log_2(p_{A,1}/q_A) & log_2(p_{A,2}/q_A) & \dots & log_2(p_{A,L}/q_A) \\ log_2(p_{C,1}/q_C) & log_2(p_{C,2}/q_C) & \dots & log_2(p_{C,L}/q_C) \\ log_2(p_{G,1}/q_G) & log_2(p_{G,2}/q_G) & \dots & log_2(p_{G,L}/q_G) \\ log_2(p_{T,1}/q_T) & log_2(p_{T,2}/q_T) & \dots & log_2(p_{T,L}/q_T) \end{bmatrix}, \tag{1.1}$$

where $p_{i,j}$ $(i \in \{A, C, G, T\}, j \in \{1, \dots, L\})$ is the frequency of nucleotide i at position j in N aligned sequences of length L, and q_i $(i \in \{A, C, G, T\})$ is the background frequency of nucleotide i. Using the PSSM matrix of a TF, we could determine if a given sequence is likely to be bound by the TF.

The TF binding motif could also be represented by motif logos graphically (Figure 1.6). The height of each position in the sequence logo is measured with the unit bit, and is equal to the Kullback–Leibler divergence (KL-divergence) between the probability distribution of nucleotides at position j and the probability distribution $q_i = 0.25$ $(i \in \{A, C, G, T\})$:

$$D_{KL}(P_j \parallel Q) = \sum_{i \in \{A,C,G,T\}} p_{i,j} log_2 \frac{p_{i,j}}{q_i}. \tag{1.2}$$

The relative height of each nucleotide in the stack is proportional to its probability at the corresponding position.

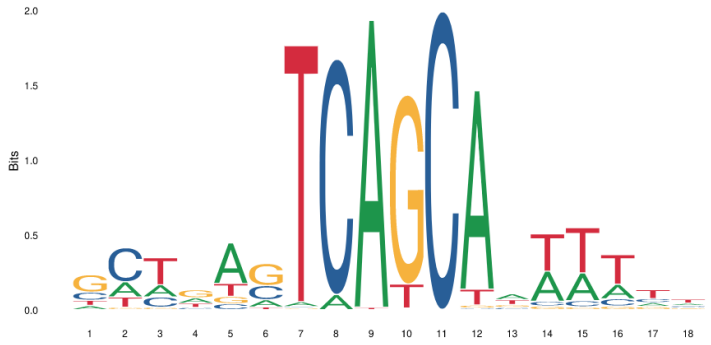


Figure 1.6: Motif logo of transcription factor MAFF (MA0495.1) from the JASPAR database [24]. The Y-axis is in bits. The core binding motif (consensus sequence) is TCAGCA or its reverse complement TGCTGA.

From the above introduction we could see that a single nucleotide change in a DNA sequence fragment might perturb the binding affinity of corresponding transcription factors, and subsequently affect the expression of genes regulated by the TFs. A single nucleotide variation may also alter the chromatin topology. Z. Tang *et al.* demonstrated that a single nucleotide variation disturbed the binding affinity of CTCF and altered the CTCF-mediated chromatin looping [25].

1.1.5 Functional analysis of LGG GWAS variants

As mentioned in Section 1.1.2, recent large-scale genome-wide association studies have implicated at least 25 SNPs associated with the LGG susceptibility, but their molecular pathways remain elusive. Only few studies have hitherto discovered the corresponding genes directly regulated by these SNPs [26, 27]. This work is dedicated to find the causative consequences of the LGG GWAS SNPs, specifically, to decipher how a single nucleotide change might contribute to the glioma phenotype.

Most of the LGG GWAS SNPs reside in the non-coding regions of the human genome, posing significant challenges to revealing their molecular functions and identifying susceptibility genes that may inform preventive and therapeutic measures. Our main hypothesis is that the GWAS loci contain causal SNPs that reside in the functional regulatory regions of the human genome and modulate the expression of target genes by directly perturbing the binding affinity of TFs. To better understand the molecular functions of germline variants in modulating LGG risk, we developed an integrative framework utilizing genomic, epigenomic and transcriptomic data to identify candidate (causal SNP, target gene, transcription factor) triplets. The steps are summarized as below, and the detailed integrative analysis pipeline for analyzing LGG GWAS SNPs is presented in Chapter 2.

We first integrated genomic and epigenomic data from public databases, such as Encyclopedia of DNA Elements (ENCODE) [17] and Roadmap Epigenomics Mapping Consortium (REMC) [28, 29], to identify germline variants located in regulatory elements of the non-coding genome. We then performed expression quantitative trait loci (eQTL) and phased allele specific expression (ASE) analyses on heterogeneous low-grade glioma data from The Cancer Genome Atlas (TCGA) [30] to prioritize putative target genes of a given GWAS SNP. At last, we performed *in-silico* TF binding sequence perturbation analysis and TF-target gene expression correlation analysis, which implicated the transcription factors whose binding affinity might be perturbed by putative causal SNPs. For the case study of 11q23.2 variant rs648044, experimental validation was also performed.

We hope our proposed (causal SNP, target gene, transcription factor) triplets could facilitate additional analysis or experimental validation. We also hope the integrative and systematic analysis of the LGG GWAS

loci could help accelerate the discovery of molecular mechanisms underlying genetic risk factors for gliomas.

1.2 Deep learning applied to transcription factor binding prediction

1.2.1 Convolutional neural network (CNN)

Convolutional neural network (CNN) has gained tremendous success in computer vision related fields. It has shown prodigious ability in pattern recognition, class identification, image segmentation etc. An example of a convolutional neural network is shown in Figure 1.7.

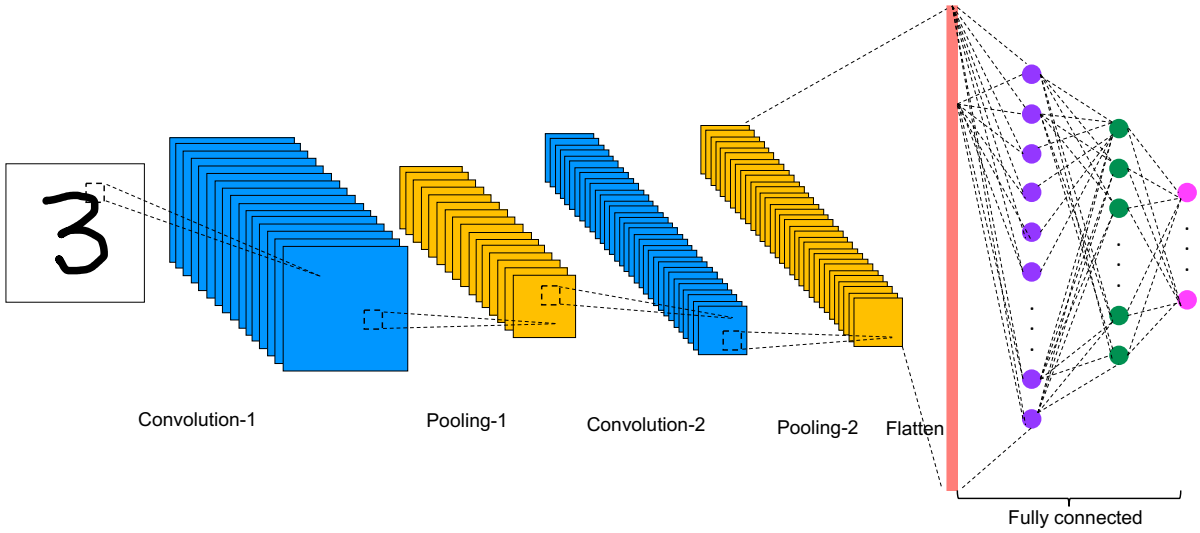


Figure 1.7: An example of a convolutional neural network.

A convolutional neural network typically includes a series of convolutional layers and pooling layers, followed by the fully connected layers (Figure 1.7). The convolutional layer is the core structure of a convolutional neural network. It consists of a set of convolutional filters which slide on the input signal. In the forward pass, each filter is convolved with the input signal, yielding the dot product of the elements of the filters with the input elements, and produce an activation map from the input. There are two commonly used activation functions: Sigmoid function, expressed as $\sigma(z) = \frac{1}{1+e^{-z}}$; ReLU function, expressed as $\sigma(z) = max(0, z)$ (Figure 1.8B). ReLU activition is usually used after the convolutional layers in a convolutional neural network, while the sigmoid function is often used at the output layer. Between two successive convolutional layers, we typically insert a pooling layer to down sample the results obtained from the convolutional layer. Max pooling is the most commonly used pooling function. It segregates the input

of the pooling layer into non-overlapping rectangular areas, and outputs the maximum number in each rectangular area. After a set of convolutional and pooling layers, the fully connected layers are appended. The fully connected layers in a CNN are essentially a "vanilla" artificial neural network (Figure 1.8A). The fully connected layer transforms a high-dimensional input signal (denoted by $\mathbf{x}$) to a high-dimensional output signal (denoted by $\mathbf{y}$) with a large dense matrix $\mathbf{W}$, bias vector $\mathbf{b}$ and an activation function σ: $\mathbf{y} = \sigma(\mathbf{W}\mathbf{x} + \mathbf{b})$.

After choosing the loss function, stochastic gradient descent (SGD) or other optimization methods modified from SGD are utilized to train a convolutional neural network. For our CNN models (Chapter 4, Chapter 6), we used the adaptive moment estimation (Adam) optimizer [31].

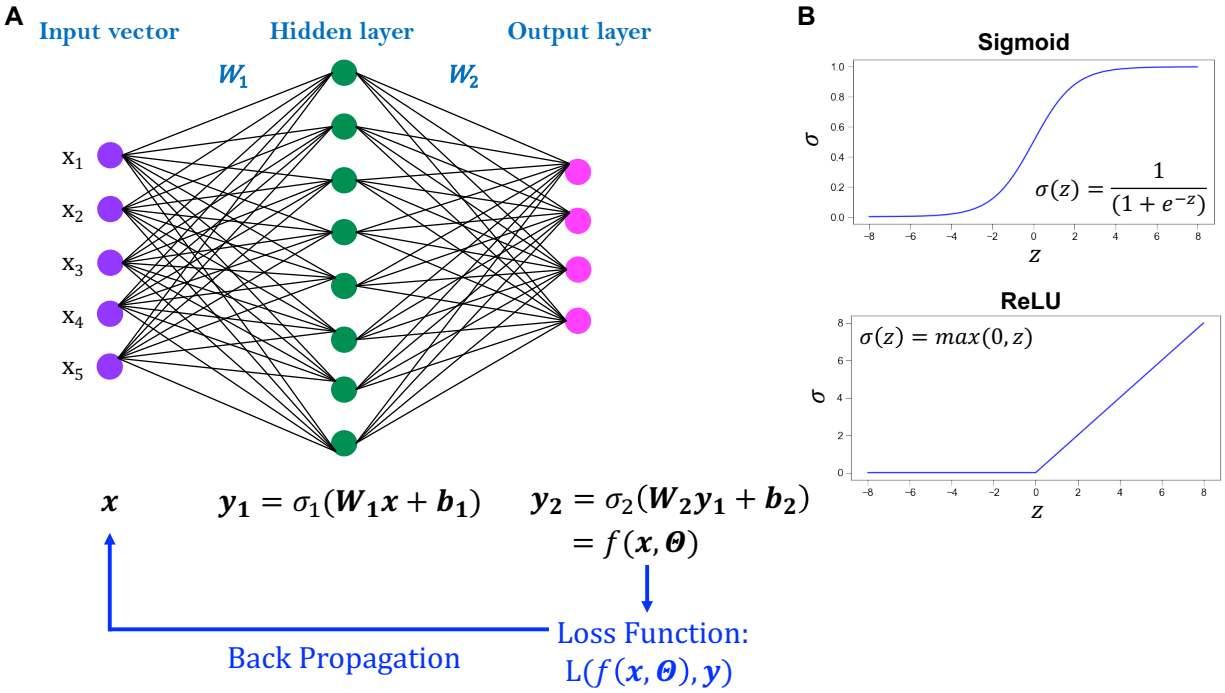


Figure 1.8: The basic structure of a feed forward neural network (A) and two commonly used activation functions: Sigmoid and ReLU (B).

1.2.2 Deep learning methods for TF binding prediction

As mentioned in Section 1.1.4, transcription factors are proteins that bind to DNA sequences and regulate gene transcription. A transcription factor recognizes a binding motif, and its binding sites are cell-type specific due to the different landscape of different cells' open chromatin regions. Experimentally, transcription factor binding sites could be determined by ChIP-seq. Despite a large number of ChIP-seq experiments have been performed, however, it is still not possible to obtain the ChIP-seq results for every transcription factor

in every cell type. Thus, predicting cell-type specific transcription factor binding sites has become a challenge and a research hotspot in bioinformatics.

Like other machine learning methods, deep learning methods have also been applied to transcription factor binding prediction. Many previous deep learning approaches have been using sequence information only, for example, the DeepBind model (Figure 1.9) [32]. Some recent studies have begun to utilize other genomic and epigenomic information in the prediction, for example, the FactorNet model (Figure 1.10) [33]. In our deep learning model (Figure 2.5) to be detailed in Section 2.3.3 and Section 4.3, we integrated DNase-seq signal with sequence information into one convolutional filter, and utilized the same set of convolutional filters to scan the positive strand sequences as well as their reverse complements (Figure 2.5). We successfully performed the allele-specific binding prediction for transcription factor SP1.

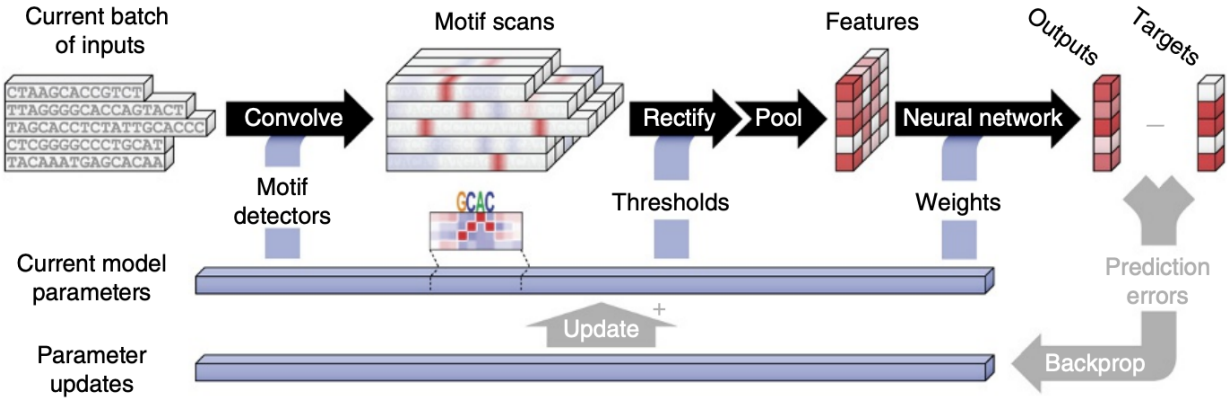


Figure 1.9: The DeepBind model, where the inputs are sequences only. Figure from B. Alipanahi *et al.* [32].

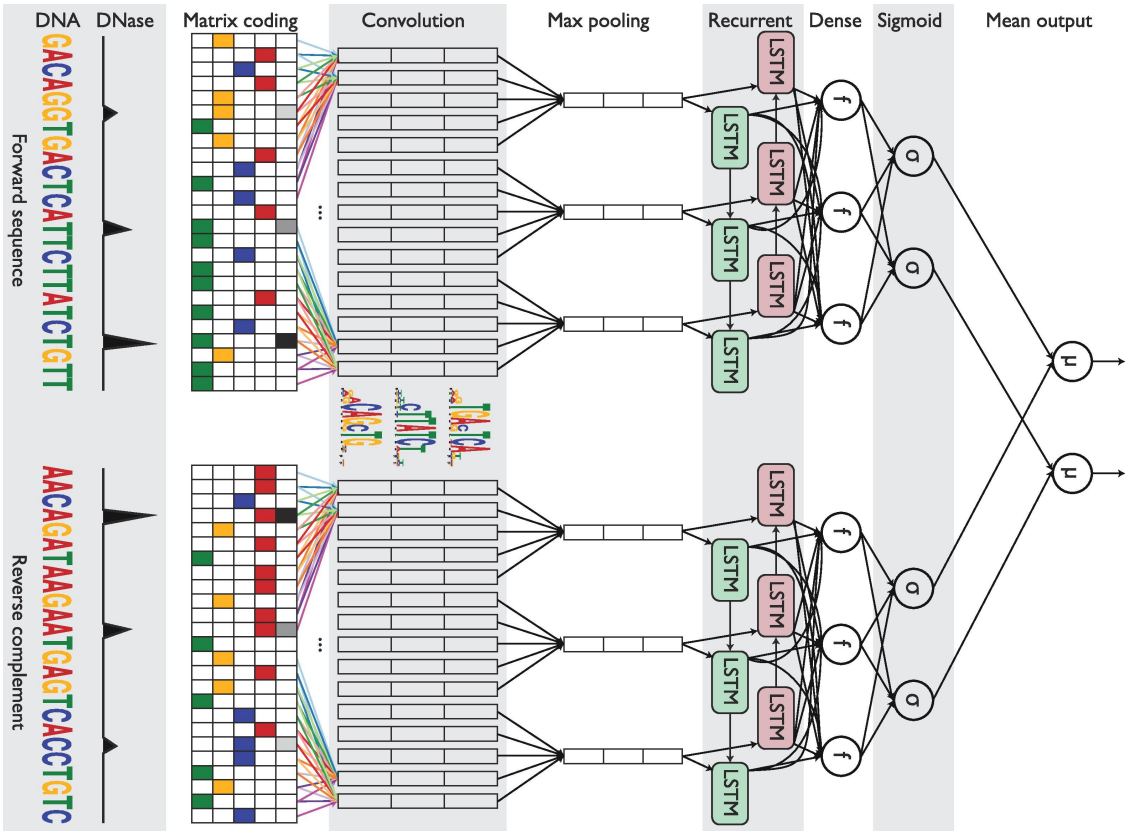


Figure 1.10: The FactorNet model, where the inputs include sequence information and open chromatin information (DNase). Figure from D. Quang *et al.* [33].

1.3 Basics of tensor and tensor decomposition

Tensors can be viewed as multidimensional generalizations of vectors and matrices. Mathematically, tensor is defined as a d-dimensional array: $\mathbf{A} = A(i_1, \ldots, i_d)$, where $\mathbf{A}$ denotes a tensor and i_k ($k = 1, \ldots, d$; $1 \leqslant i_k \leqslant n_k$) denotes the index of dimension. Tensor are widely used in various areas. When performing tensor analysis, however, we often face the problem of *curse of dimensionality*, where memory and amount of operations grows exponentially in dimension d [34]. To work with large dimensional problems, an efficient representation of a tensor by a small number of parameters is thus needed.

There are different ways for performing tensor decomposition, for example, canonical polyadic decomposition (CP decomposition) and higher-order singular value decomposition (HOSVD). Canonical polyadic decomposition is also called tensor rank decomposition, and is a generalization of the matrix singular value

decomposition (SVD) to tensors. Given a tensor $\mathbf{A}$, CP decomposition is expressed as:

$$A(i_1, \ldots, i_d) = \sum_{\alpha=1}^{r} U_1(i_1, \alpha) U_2(i_2, \alpha) \ldots U_d(i_d, \alpha), \tag{1.3}$$

where r is the tensor rank, and represents the minimal summands to express $\mathbf{A}$ in equation 1.3; $U_k = [U_k(i_k, \alpha)]$ are called canonical factors [34]. Equation 1.3 is a good candidate for expressing high dimensional tensor $\mathbf{A}$ in low-parametric format. However, there are several drawbacks when numerically computing such an approximate representation: first, computing the tensor rank is an NP-hard problem [35, 34]; second, the approximation of $\mathbf{A}$ in the Frobenius norm with a fixed tensor rank could be ill-posed [36, 34]; moreover, there is a probability of being stuck in the local minima in the computation process [34]. Thus, it is numerically difficult to get the best low-tensor-rank approximation of $\mathbf{A}$. On the other hand, higher-order singular value decomposition (HOSVD), which is an orthogonal format of Tucker decomposition [37, 38], provides a way to decompose a tensor into a small core tensor and several matrices. Higher-order SVD of a given tensor $\mathbf{A}$ is written as:

$$\mathbf{A} = \mathbf{S} \times_{\mathbf{1}} \mathbf{U^{(1)}} \times_{\mathbf{2}} \mathbf{U^{(2)}} \ldots \times_{\mathbf{d}} \mathbf{U^{(d)}}, \tag{1.4}$$

where $\mathbf{U^{(k)}} = (U_1^{(k)}, \ldots, U_{n_k}^{(k)})$ is an $n_k \times n_k$ orthogonal matrix, and $\mathbf{S}$ is an "all-orthogonal" and ordered core tensor. The truncated HOSVD of $\mathbf{A}$, written as $\hat{\mathbf{A}}$, is then obtained by discarding the first few left singular vectors with the smallest singular values in HOSVD [39]. There are also several disadvantages of higher-order SVD or Tucker decomposition: although the Tucker format is stable, its parameter number $O(dnr + r^d)$ depends on the exponent d, and thus is not suitable for large dimension cases [34]; unlike in the matrix case, when performing tensor approximation using truncated HOSVD, the truncated approximation may not be the best possible approximation that satisfies the n-mode rank [1] constraints [39].

All the above disadvantages of CP decomposition and the Tucker format motivate us to study a new format of tensor decomposition, called tensor train decomposition, abbreviated as TT-decomposition. It is also known as matrix product state (MPS) in theoretical condensed matter physics. A tensor $\mathbf{A}$ is said to be in the tensor train format (TT-format) if element with index $(i_1, \ldots, i_d)$ in tensor $\mathbf{A}$ is expressed as [34]:

$$A(i_1, \ldots, i_d) = G_1(i_1) G_2(i_2) \ldots G_d(i_d), \tag{1.5}$$

[1] n-mode rank (from [39]): Suppose the HOSVD of tensor $\mathbf{A}$ is expressed as $\mathbf{A} = \mathbf{S} \times_{\mathbf{1}} \mathbf{U^{(1)}} \times_{\mathbf{2}} \mathbf{U^{(2)}} \ldots \times_{\mathbf{d}} \mathbf{U^{(d)}}$, and $A_{(n)}$ is the mode-n unfolding of tensor $\mathbf{A}$. Because the matrices $\mathbf{U^{(n)}}$ are orthogonal, $\|\mathbf{A}\|_F^2 = \|\mathbf{S}\|_F^2 = \sum_{i=1}^{R_1} (\sigma_i^{(1)})^2 = \ldots = \sum_{i=1}^{R_d} (\sigma_i^{(d)})^2$, where R_n, called the n-mode rank or n-rank, is the matrix rank of $A_{(n)}$.

where $G_k(i_k)$ is an $r_{k-1} \times r_k$ matrix, and $r_0 = r_d = 1$ [34]. It also can be written in the index form [34]:

$$A(i_1, \ldots, i_d) = \sum_{\beta_0 \ldots \beta_d} G_1(\beta_0, i_1, \beta_1) G_2(\beta_1, i_2, \beta_2) \ldots G_d(\beta_{d-1}, i_d, \beta_d). \tag{1.6}$$

We will give the details of the TT-format of a given tensor in Chapter 6. Representing a tensor using the TT-format is referred to as TT-decomposition.

There have been a number of applications of TT-decomposition. In Chapter 6, we will first review the idea of applying TT-decomposition to neural network parameter reduction in the fully connected layer, proposed by A. Novikov *et al.* [40], and then demonstrate the performance of the parameter-reduced CNN of SP1. Similar to the idea of factorizing the fully connected layer [40], Y. Yang *et al.* applied tensor train recurrent neural network to video classification [41]. In the work, the authors utilized TT-decomposition to reduce the parameter number of the input-to-hidden mapping in the recurrent neural network model, and demonstrated its competitive prediction accuracy with the state-of-art models in video classification. J. Su *et al.* proposed a convolutional tensor-train decomposition model (CTTD), where convolutional kernels were factorized into tensors, and applied CTTD to construct a convolutional tensor-train long short-term memory (LSTM) model to capture long-term spatio-temporal correlations [42].

Tensors are not only widely used in various areas of computer sciences, but also heavily used in theoretical condensed matter physics. For example, we could use tensor network methods to simulate strongly correlated systems, and describe the wave function of the system as a network of tensors [43, 44, 45, 46]. Density matrix renormalization group (DMRG) [47, 48, 49, 50] is another famous example of a tensor network method [46].

1.4 Overview

This work is dedicated to understand the molecular functions of germline genetic variants in modulating LGG risk using statistics and physics-inspired deep learning methods. The thesis contains three interconnected modules. In the first module, we present an integrative computational framework to identify the putative (causal SNP, target gene, transcription factor) triplets that might contribute to LGG oncogenesis (Chapter 2). As case studies of detailed analysis and interpretation, we then focus on 3 loci that have (1) a target gene with known tumor suppressor functions in other cancers (rs648044, *ZBTB16* locus, Chapter 3), (2) one of the lowest GWAS *P*-values (rs12803321, *PHLDB1* locus, Chapter 4), and (3) no convincing eQTL candidate gene in a previous study [12] (rs11706832, *LRIG1* locus, Chapter 5), respectively. To better assess the allele-specific binding pattern of the TF SP1 in the human brain, in the second module, we present a deep learning approach based on convolutional neural network and simulated-annealing-based interpretation

method (Chapter 4, Section 4.3). Inspired by the concept of tensor train decomposition, also known as matrix product state in condensed matter physics, in the third module, we present the recent proposed application of TT-decomposition to neural network parameter reduction, and demonstrate the performance of the parameter-reduced convolutional neural network of SP1 (Chapter 6). We conclude this thesis with Chapter 7, conclusion.

Our main results and methods from Chapter 2, 3, 4, 5 have been published in Neuro-oncology as a research article: Manjunath[†], Yan[†] *et al.*, "Functional analysis of low-grade glioma genetic variants predicts key target genes and transcription factors" ([†]co-first authors, contributed equally) [51]. The author of this thesis also contributed to the LGG-related analysis in Zhang *et al.*, "The cancer-associated genetic variant rs3903072 modulates immune cells in the tumor microenvironment" [52], where the author performed eQTL analysis in LGG and confirmed the negative association between the expression level of *CTSW* and the risk genotype of a breast cancer GWAS SNP rs3903072 in low grade glioma. This helped to confirm the putative role of rs3903072 in modulating the expression of *CTSW*, a candidate tumor suppressor, in tumor-infiltrating immune cells [52]. Due to the focus of this thesis, we will not expand on the work done by the author in the above-mentioned *CTSW* paper.

Chapter 2

Integrative analysis pipeline for LGG GWAS SNPs

In this chapter, we present an integrative analysis pipeline for LGG GWAS loci. Most of the work presented in this chapter has been published in Manjunath†, Yan† [1] *et al.*, Neuro-oncology, 2020 [51]. The computational part of the pipeline was developed in collaboration with Dr. Mohith Manjunath and Dr. Yi Zhang. The experimental parts were performed by members of Professor Paul R. Selvin's lab at UIUC, members of Professor Robert B. Jenkins's lab at Mayo Clinic and members of Professor Joseph Costello's lab at UCSF. To understand the functional impact of LGG GWAS variants, this framework integrated heterogeneous genomic, epigenomic and transcriptomic high-throughput sequencing data, and incorporated computational, experimental and deep learning approaches (Figure 2.1). The pipeline could be divided into three subsections: putative causal SNP identification, target gene identification, and TF prioritization. We started with a list of 25 GWAS SNPs that showed association with increased risk for LGG in the population (Table A.1, Appendix A.1).

2.1 Putative causal SNP identification

The genome-wide association studies usually report the SNPs that are most significantly associated with some trait (smallest P-values in GWAS) as the GWAS SNPs. However, the reported GWAS SNPs might not necessarily be the functionally causal ones. We proposed that the nearby SNPs residing in open regulatory chromatin regions and in high linkage disequilibrium (LD) with the GWAS SNPs could act as true molecular effectors. We therefore examined all SNPs in high LD with the GWAS SNPs ($r^2 \geqslant 0.8$, 1000 Genomes Phase 3, EUR population, Appendix A.1) along with the epigenomic information to identify the putative causal ones.

[1] †co-first authors, contributed equally.

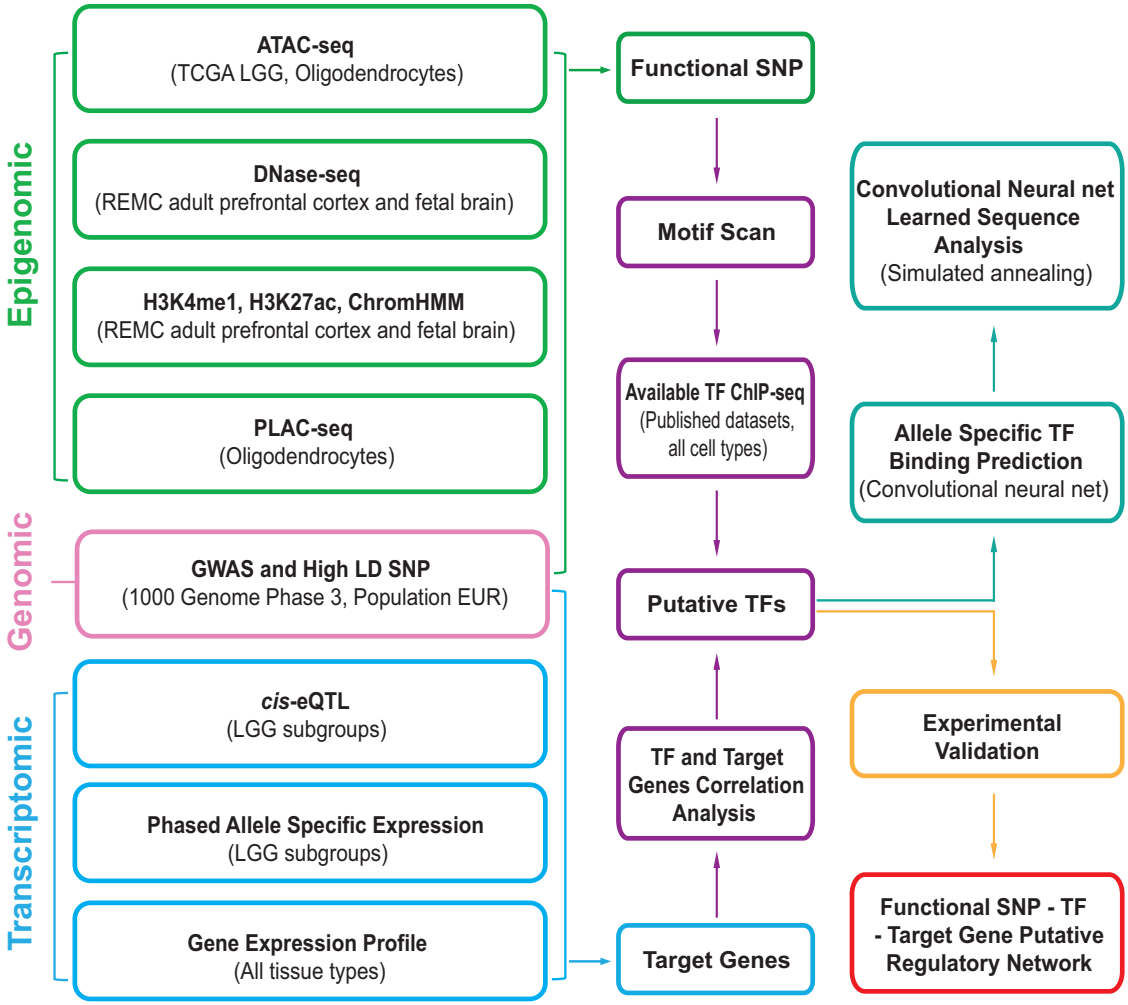


Figure 2.1: Integrated framework for functional analysis of LGG GWAS SNPs. Green: epigenomic data; pink: genomic information; blue: transcriptomic data and analysis; purple: motif and TF-gene expression correlation analyses; ocean blue: deep learning approaches for TF binding prediction; yellow: experimental validation; red: candidate triplets.

2.1.1 Epigenomic data

Epigenetic modifications, such as DNA methylation and histone modification, are reversible modifications on DNA or histone without altering the sequence of DNA. They play an important role in modulating gene expression and are involved in numerous cellular processes including tumorigenesis. Histone H3 lysine 4 mono-methylation (H3K4me1) and H3K27 acetylation (H3K27ac) are enriched at active enhancers, while H3K4 tri-methylation (H3K4me3) is enriched at promoters. In order to select high LD SNPs residing in regulatory enhancer regions of the brain genome, we examined H3K4me1 and H3K27ac chromatin immunoprecipitation followed by sequencing (ChIP-seq) data from the fetal brain and dorsolateral prefrontal cortex. Trained on a set of histone marks, ChromHMM [53] could capture the combinatorial interactions between different histone marks; the trained model could then be used to calculate the posterior probability of each chromatin state for different genomic regions. We therefore additionally utilized the processed ChromHMM data tracks in our analysis.

We also assessed the open chromatin information given by assay for transposase-accessible chromatin using sequencing (ATAC-seq) and DNase I hypersensitive sites sequencing (DNase-seq) as well as chromatin interactions measured by proximity ligation-assisted ChIP-seq (PLAC-seq), to further prioritize the high LD SNPs residing in accessible regulatory DNA elements in the human brain.

The epigenomic data sources are stated as below. We obtained histone modification (H3K4me1, H3K27ac) ChIP-seq datasets of brain tissues, DNase-seq datasets of the fetal brain and ChromHMM [53] tracks of different tissues from the Roadmap Epigenomics Mapping Consortium (REMC) database [28, 29]. Primary tumor ATAC-seq aligned BAM files of glioma patients were downloaded from the TCGA Data Portal [30], and the normalized ATAC-seq BigWig files of glioma patients were obtained from Corces *et al.* [54]. The processed ATAC-seq, H3K27ac and PLAC-seq data in oligodendrocytes were obtained from Nott *et al.* [55] (https://genome.ucsc.edu/s/nottalexi/glassLab_BrainCellTypes_hg19).

2.1.2 Linkage disequilibrium and candidate causal SNPs

Linkage disequilibrium (LD) describes the non-random association of the alleles at two loci in one population [56, 57]. The level of linkage disequilibrium is affected by natural selection, genetic drift, mutation, recombination etc. [57] One of the measures used in the community to describe LD between two loci is r squared (r^2, $0 \leqslant r^2 \leqslant 1$), where larger r^2 denotes higher linkage disequilibrium. In human genetics, if two SNPs are in high LD, the allele frequencies of these two SNPs are strongly associated, as illustrated by an example in Figure 2.2. Thus, if one SNP is reported as GWAS SNP in genome-wide association studies, its high LD SNPs are also likely significantly associated with the phenotype, although with slightly larger

P-value compared to the GWAS SNP (Figure 2.3).

We therefore proposed that the nearby SNPs which reside in open regulatory chromatin regions and in high linkage disequilibrium with the GWAS SNPs could act as true molecular effectors. We obtained all SNPs in high LD with the GWAS SNPs ($r^2 \geqslant 0.8$, 1000 Genomes Phase 3, EUR population, Appendix A.1) using LDlink [58]. We then aligned all the high LD SNPs with the DNase-seq, ATAC-seq, PLAC-seq H3K4me1 ChIP-seq and H3K27ac ChIP-seq signal tracks, and selected the SNPs that reside in the peaks of the above-mentioned signals as putative causal SNPs.

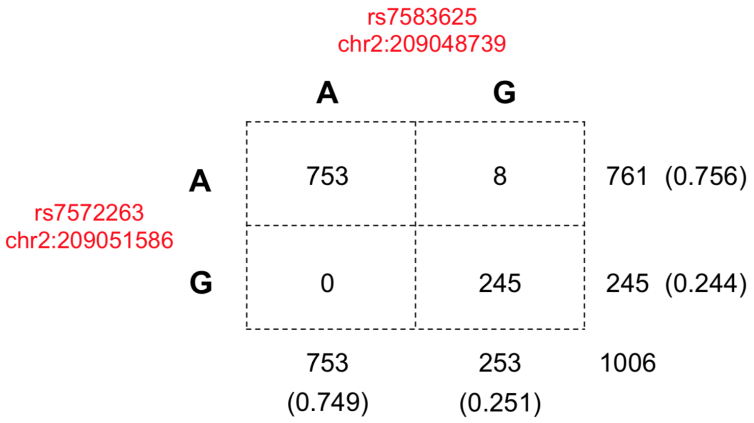


Figure 2.2: A snapshot given by LDlink [58] of two high LD SNPs in EUR population. Haplotype counts: rs7572263-A|rs7583625-A: 753; rs7572263-A|rs7583625-G: 8; rs7572263-G|rs7583625-A: 0; rs7572263-G|rs7583625-G: 245. Allele frequencies: rs7572263-A: 0.756; rs7572263-G: 0.244; rs7583625-A: 0.749; rs7583625-G: 0.251. r^2 of these two SNPs: 0.9582. rs7572263-A allele is correlated with rs7583625-A allele; rs7572263-G allele is correlated with rs7583625-G allele.

2.2 Target gene identification

We performed expression quantitative trait loci (eQTL) and phased allele specific expression (ASE) analyses using The Cancer Genome Atlas (TCGA) genotype and gene expression profiles to identify putative target genes.

2.2.1 TCGA LGG data

We utilized five types of TCGA LGG datasets [30]: germline genotype data of 513 patients, primary tumor copy number segmentation data of 513 patients, tumor RNA-Seq aligned bam files of 516 patients, processed gene-level RSEM (RNA-Seq by Expectation Maximization) expression data of 516 patients, and clinical

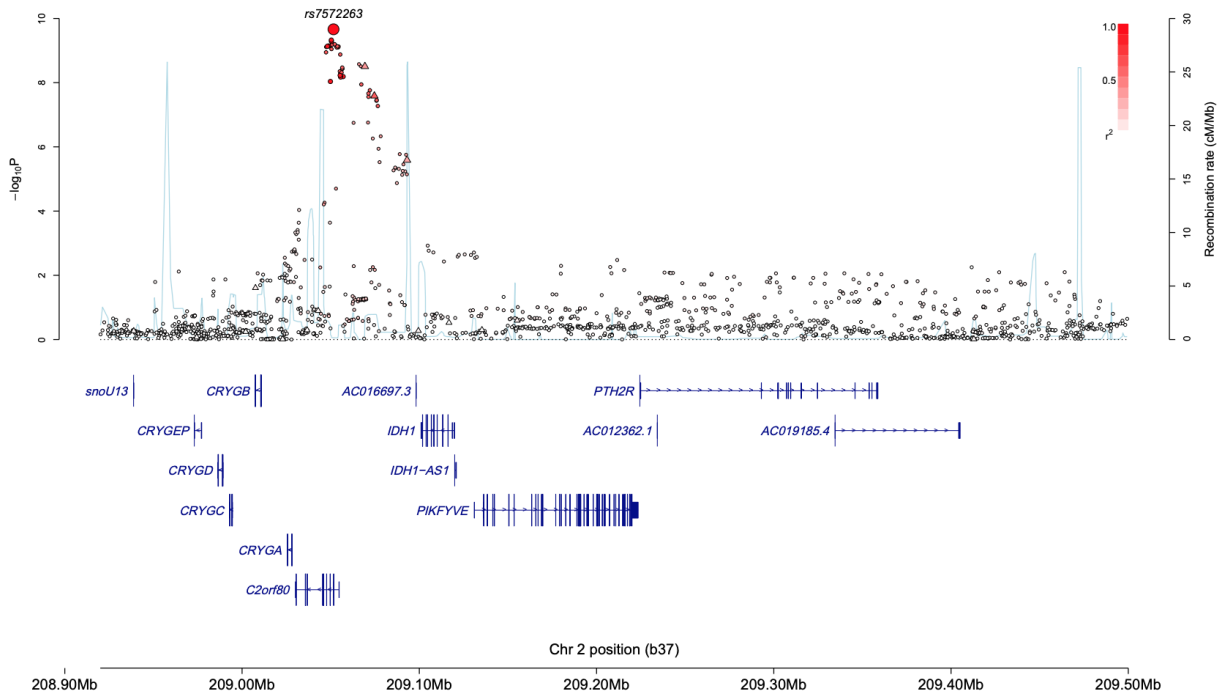


Figure 2.3: The GWAS P-value landscape in the region around LGG GWAS SNP rs7572263. This figure is Supplementary Fig. 3 of Melin *et al.*, Nat. Genet., 2017 [12], and is shown here to illustrate the P-value landscape of the reported GWAS SNP and its high LD SNPs. Y-axis on the left: $-\log_{10}(P$ value from GWAS); Y-axis on the right: genetic recombination rate, shown as light blue line in the figure; X-axis: human NCBI build 37 Chromosome 2 positions; colored triangle or circle: genotyped (triangle) or imputed (circle) SNPs with the color corresponding to the r^2 of LD with the lead SNP (rs7572263). The genes mapping to this region were also shown in the figure.

data of 515 patients. Out of 508 patients possessing all five data types, 427 patients' molecular subtype information was available [59, 60]. Assigning these 427 patients to the five molecular subgroups ("*TERT* promoter mutation only", "IDH^{mut} only", "*TERT* promoter mutant and IDH^{mut}", "triple-positive" and "triple-negative") yielded 204 patients in the "IDH^{mut} only" subgroup and 137 patients in the "triple-positive" subgroup.

2.2.2 Genotype imputation

Since genotype files from TCGA only provide the genotype calls of 906600 tag SNPs, we need to impute the genotype and haploptype of all SNPs using imputation engines. The preparation of imputation input file and the process of imputation results are stated as below.

We first obtained germline genotype data of 513 LGG patients in the birdseed format from the TCGA GDC Data Portal Legacy Archive [30]. Instead of tumor tissues, genotype data from patients' normal tissues were used for all analyses to avoid miscalls from genotyping error and somatic mutations. The genotypes

of the tag SNPs measured by Affymetrix human SNP array 6.0 were matched to hg19 coordinates using Affymetrix genome-wide SNP annotation file. Tag SNPs with genotype confidence score > 0.01 were filtered out. Untagged SNPs were imputed and phased from tagged SNPs of 513 LGG patients using the Michigan Imputation Server [61]. We chose Haplotype Reference Consortium (HRC) panel [62] (version r1.1 2016) which consists of 64940 haplotypes of predominantly European ancestry as the imputation reference panel and Eagle2 [63] as the phasing engine. After the imputation, imputed SNPs were retained if the minor allele frequency (MAF) exceeded 0.005 and estimated imputation accuracy (R^2) exceeded 0.4. Then, imputed genotypes were retained if the maximum of the estimated posterior probabilities for genotypes 0/0, 0/1 and 1/1 was larger than or equal to 0.9. Here, 0 denotes the SNP's reference allele and 1 denotes the SNP's alternative allele. We thus could extract the imputed genotypes of all SNPs. We also extracted the haplotype of SNPs using the imputed and phased results from the Michigan Imputation Server [61] (Section 2.2.4).

2.2.3 Expression quantitative trait loci (eQTL) linear model

Expression quantitative trait loci are genomic loci which explain the variation in the expression levels of genes [64, 65]. eQTL analysis is widely used in the functional genomics field to identify the target gene whose expression level, referred to as a quantitative trait, is associated with a single nucleotide polymorphism. In our study, we performed expression quantitative trait loci analysis using the TCGA LGG data to identify candidate target genes associated with the GWAS SNPs. We imputed high-confidence genotypes at the GWAS SNPs and restricted the eQTL analysis to the genes residing within the 4Mb window centered at each GWAS SNP. The choice of the 4Mb window was based on our underlying hypothesis that the SNPs function by perturbing the interactions of enhancers and promoters. Enhancer-promoter interactions tend to occur within the topologically associated domains (TADs), 92% of whose size lie between 200 and 2.5Mb [66]. Thus, we chose 4Mb window centered around each GWAS SNP for the eQTL analysis. We used the following multivariate linear regression model to assess the association between the GWAS SNP's genotype and the gene expression level, while adjusting for gene copy number, tumor site, tumor grade, histological diagnosis, and gender:

$$E_i = \alpha_i + \beta_i \cdot GT + \gamma_i \cdot \overline{CNS_i} + \sum_{j=1}^{4} \theta_{ij} Cov_j + \epsilon_i. \tag{2.1}$$

In equation 2.1, i indexes the genes within the 4Mb window centered at the SNP; $E_i = log_2(\text{RSEM} + 1)$ denotes the log-transformed gene expression level in RSEM unit; GT (genotype) $\in \{0, 1, 2\}$ denotes the number of alternative alleles of the GWAS SNP with respect to the human reference genome; $\overline{CNS_i}$ is

the length-weighted average of tumor copy number segmentation, $log_2(\frac{copy\,number}{2})$, covering gene i; Cov_j represents each of the 4 covariates included in the eQTL analysis: tumor site, tumor grade, histological diagnosis, and gender, where tumor site $\in$ {"supratentorial, frontal lobe", "supratentorial, occipital lobe", "supratentorial, parietal lobe", "supratentorial, temporal lobe", "supratentorial, not otherwise specified"}, tumor grade $\in$ {"grade II", "grade III"}, histological diagnosis $\in$ {"astrocytoma", "oligodendroglioma", "oligoastrocytoma"} and gender $\in$ {"male", "female"}; α_i denotes the intercept; and, ϵ_i denotes the error term.

2.2.4 Phased allele-specific expression (ASE) analysis

From the eQTL analysis, genes with false discovery rate (FDR) [67] adjusted $p_i \leq 0.2$, where p_i is the P-value of the genotype linear regression coefficient, were selected as candidate target genes. For each candidate gene, we performed a phased allele-specific expression (ASE) analysis to test the differential transcription between the two chromosomes harboring different alleles of a given GWAS SNP [68] (Figure 2.4). We first obtained a subset of patients having heterozygous genotypes both at the GWAS SNP and at exonic SNPs of the candidate gene. We then extracted the imputed haplotype from the imputation results (Section 2.2.2) to determine the phase between the GWAS SNP and the exonic SNPs. Allele-specific coverage of the exonic SNPs by RNA-Seq reads ($MAPQ \geq 20$) was obtained using the code from Zhang *et al.* [68] and Wilcoxon signed-rank sum test (for sample size $n \geq 5$) was used to examine the transcription imbalance between the two copies of chromosomes at a P-value threshold of 0.05.

2.3 TF prioritization

We have thus identified putative causal SNPs that reside in accessible regulatory chromatin regions (Section 2.1) and a set of putative target genes (Section 2.2). To verify our central hypothesis that the causal SNPs modulate the expression of target genes through perturbing the binding affinity of transcription factors (TFs), we further performed a series of analyses to prioritize candidate TFs. We first performed motif analyses to select a set of TFs whose DNA binding affinity were significantly perturbed by the single nucleotide change of candidate causal SNPs [68] (Section 2.3.1). We further filtered candidate TFs based on TF-target gene correlation analysis (Section 2.3.2). To assess the impact of causal SNPs on candidate TFs, we at last trained a convolutional neural network model on TF ChIP-seq data to predict allele-specific TF binding (Section 2.3.3) and deployed a simulated annealing method [69, 70] to extract the optimal TF motif learned by the CNN (Section 2.3.4). The TF motif analyses (Section 2.3.1) were developed by Dr. Yi Zhang and Dr.

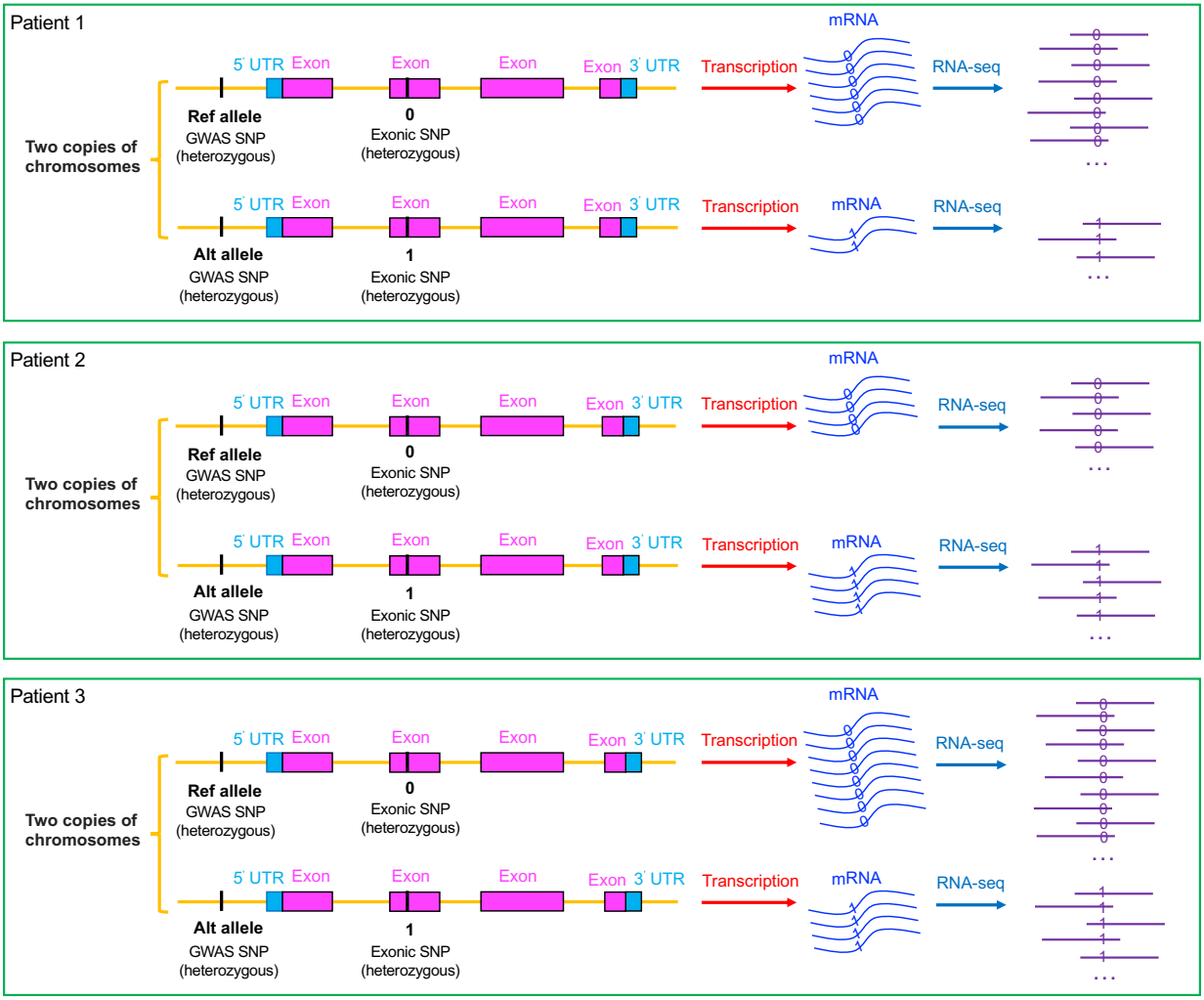


Figure 2.4: Transcription imbalance from two copies of chromosomes could be revealed by phased ASE analysis. Illustrative plots of 3 patients are shown in the figure. Heterozygous GWAS SNP and exonic SNP are shown as small black bars in the figure. "0" denotes exonic SNP's allele that in the same chromosome (in phase) with the GWAS reference allele; "1" denotes exonic SNP's allele that in the same chromosome (in phase) with the GWAS alternative allele. Transcription imbalance from two copies of chromosomes is represented by the amount of mRNA transcripts; "0" or "1" in mRNA represents the chromosome from which the transcripts are transcribed. Through RNA-Seq, phased allele-specific coverage of the exonic SNP could be obtained, and transcription imbalance could then be assessed for a pool of patients (samples) using Wilcoxon signed-rank sum test (for sample size $n \geq 5$).

Mohith Manjunath [68]. I will rephrase the methods briefly in Section 2.3.1 and Appendix A.2. Dr. Mohith Manjunath also performed the ATAC-seq allele-specific read coverage analysis described in Section 2.3.3.

2.3.1 Motif analyses

We utilized motif scan software Find Individual Motif Occurrences (FIMO) [71] and 4 transcription factor motif databases JASPAR [72], HOCOMOCOv10 [73], TRANSFAC [74], Jolma2013 [75] to select TFs whose

motifs matched the flanking sequences covering causal SNPs [68]. FIMO computes the score for the match of a position in a sequence to a motif given a position-specific scoring matrix (PSSM) which represents the motif, and converts this score to a P-value, which is defined as the probability of a random subsequence of the length of the motif scoring better than or as well as the original match [71]. We scanned $\pm$ 25bp flanking sequences around putative causal SNPs (total length 51bp), and used P-value 10^{-3}, 10^{-4} as FIMO output thresholds to select a set of candidate TFs for downstream motif analysis [68].

We then performed motif permutation test using the output from FIMO [68]. The motif permutation test aims to assess the significance of the motif disruption by the causal SNP through simulating neutral mutations of the motif-matching sequence [68]. The details of motif permutation test are rephrased in Appendix A.2. Setting permutation test P-value threshold as 0.05, we thus selected all TFs whose binding affinity were significantly perturbed by the SNP.

2.3.2 TF-target gene correlation analysis

We then performed TF-target gene correlation analysis to further prioritize candidate TFs obtained from the motif analyses. We started with a set of target genes from eQTL analysis and a set of TFs from the motif analyses. We calculated Pearson correlation coefficients and Spearman correlation coefficients between the target genes' and TFs' expression values (log transformed RSEM values, $log_2(\text{RSEM}+1)$) stratified into three genotype groups (homozygous risk, heterozygous, homozygous non-risk). We performed the TF-target gene correlation analysis both in LGG subgroups ("IDH^{mut} only", "triple-positive") and in the combined groups (all 5 molecular groups combined, "IDH^{mut} only" and "triple-positive" combined). We then selected the TFs based on the reasoning that the correlation or anti-correlation would likely be strongest in the homozygous genotype group creating the TF binding motif, while the correlation would likely be weakest in the homozygous genotype group disrupting the TF binding motif.

2.3.3 Allele-specific TF binding prediction using convolutional neural network

Based on the motif analyses and TF-target gene correlation analysis, we selected a set of candidate TFs satisfying the following criteria: 1. the binding affinity of the TF was significantly perturbed by the candidate causal SNP (motif permutation test P-value < 0.05); 2. the TF-target gene correlation stratified into three genotype groups (denoted as AA, AB, BB) satisfied one of the following conditions in Table 2.1.

We further conducted two additional steps to identify the best candidate TF(s) out of the selected ones. First, we tried to infer the chromosomal accessibility difference caused by the SNP through ATAC-seq allele-specific read counts analysis. We extracted the read counts by allele from ATAC-seq aligned reads of

Target gene expression (E)	Allele harbored by the binding motif	TF-target gene correlation (r)	TF role
$E_{AA} > E_{AB} > E_{BB}$	A	$r_{AA} > r_{AB} > r_{BB}$, $r_{AA} > 0$ and strongest	Activator
$E_{AA} > E_{AB} > E_{BB}$	B	$r_{AA} > r_{AB} > r_{BB}$, $r_{BB} < 0$ and strongest	Repressor
$E_{AA} < E_{AB} < E_{BB}$	A	$r_{AA} < r_{AB} < r_{BB}$, $r_{AA} < 0$ and strongest	Repressor
$E_{AA} < E_{AB} < E_{BB}$	B	$r_{AA} < r_{AB} < r_{BB}$, $r_{BB} > 0$ and strongest	Activator

Table 2.1: A summary table of target gene expression, allele harbored by the binding motif, TF-target gene correlation coefficient and the role of TF.

heterozygous TCGA-LGG samples at the causal SNP using bcftools mpileup [76] option. We considered only the bases with a Phred quality score ≥ 20. For each sample, the significance of the skew between the two alleles was evaluated using a binomial test. The resulted P-values were combined using the Fisher's method. If a significant skew toward the risk allele was observed, the risk allele might function by creating a TF binding motif; whereas if a significant skew toward the non-risk allele was observed, the risk allele might function by disrupting a binding motif, and the non-risk allele might function by preserving a binding motif. Second, we examined candidate TFs' all available ChIP-seq datasets from public databases, and performed ChIP-seq allele-specific read counts analysis. We checked if the causal SNP resided in the peaks of the TF ChIP-seq signal; we also examined if there was a significant skew between the read counts of the two alleles in heterozygous cell lines [68]. The significance of a skew was measured by a binomial test, and one-sided P-value was reported.

However, very few TF ChIP-seq datasets were available for brain-related cell types in public databases such as Encyclopedia of DNA Elements [17] (ENCODE). Therefore, we could not verify the allele-specific binding of some candidate TFs in brain cell types. Thus, for the candidate TF SP1 in rs12803321 locus (Chapter 4), we trained a convolutional neural network (CNN) model on DNase-seq data and TF ChIP-seq data of available cell types from ENCODE, and used the model to predict the allele-specific binding pattern of the same TF in the human fetal brain.

We trained the CNN using the Keras package [77] and engineered several layers using TensorFlow [78]. The structure of the constructed CNN is shown in Figure 2.5. Each input sample consisted of a 1001×9 matrix containing the one-hot encoded forward and reverse DNA sequence information and the quantile-normalized DNase-seq signal in the given 1001bp region. The CNN model had one convolutional layer consisting of 40 convolutional filters. All convolutional filters had a size of 12×5 and slid on a 1001×5 input matrix representing positive strand DNA sequence and DNase-seq signal. To capture the motifs present in

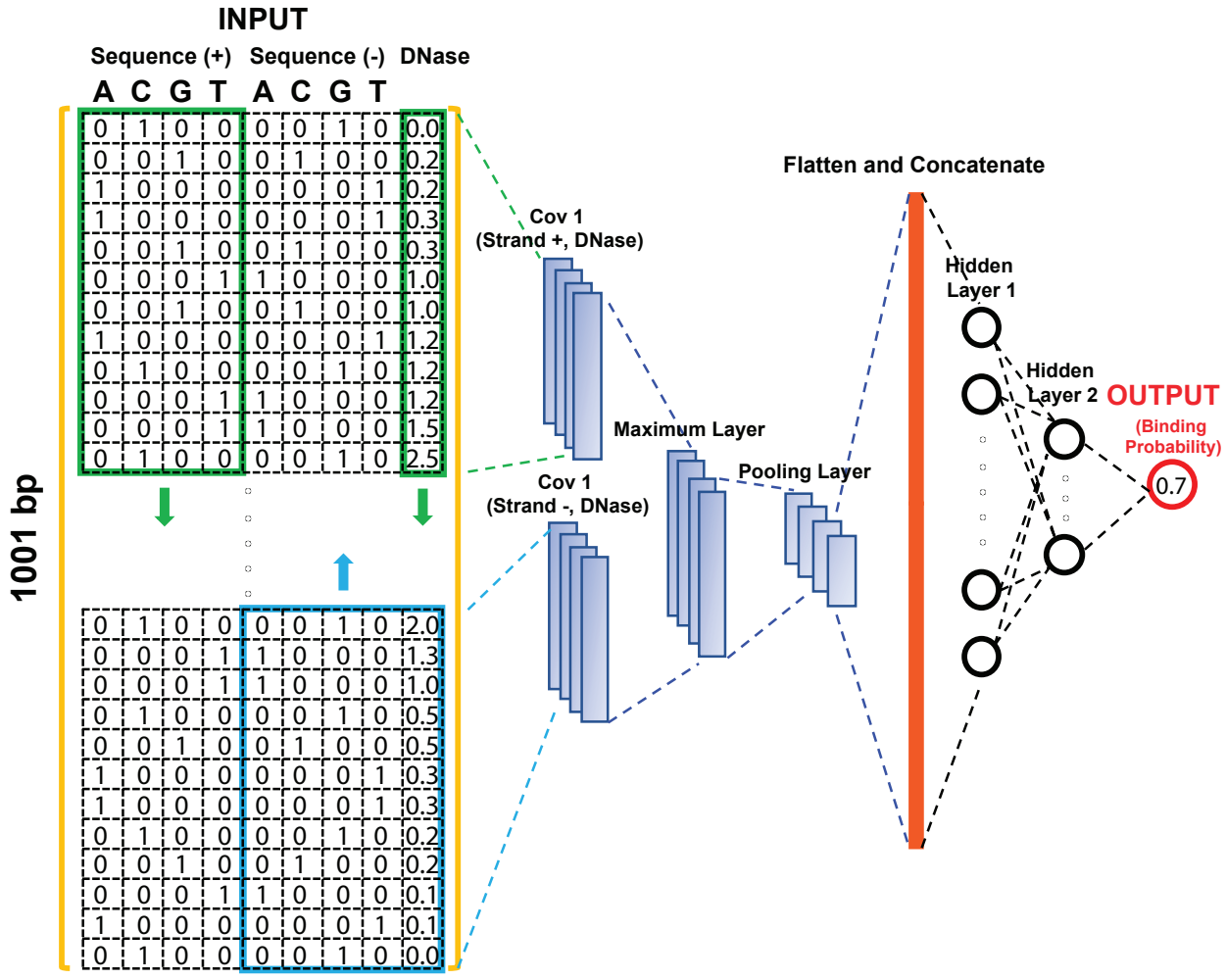


Figure 2.5: CNN for predicting the binding pattern of transcription factor SP1 based on DNA sequence and open chromatin information. From left to right: 1001bp×9 input matrix incorporating sequence information and quantile-normalized DNase-seq signal at each base; convolutional layer using filters of length 12bp; maximum layer, extracting the maximum of the convolutional layer output from the positive and negative strands; maximum pooling layer; flatten and concatenate layer; fully connected layer with 80 neurons; fully connected layer with 10 neurons; output.

negative strands, the 1001×5 submatrix representing the corresponding reverse complement DNA sequence and DNase-seq signal was passed through the same set of convolutional filters. We then extracted the maximum of the convolutional layer output from the positive and negative strands, and passed the output through a max pooling layer of size 40 and stride 40. The max pooling layer output was flattened and passed to a fully connected layer of 80 neurons, and then was passed through a fully connected layer of 10 neurons. Finally, the output of the second fully connected layer was passed to a single output neuron encoding the

binding probability of the TF. Rectified linear unit (ReLU) function was used as the activation function throughout the CNN, except for the output layer where we utilized a sigmoid function to restrict the output between 0 and 1. The loss function of the CNN model is Binary Cross-Entropy (CE) loss:

$$CE = \frac{1}{N}\sum_{i=1}^{N}[-t_i log(\sigma_i) - (1-t_i) log(1-\sigma_i)], \tag{2.2}$$

where N is the batch size; $t_i \in \{0, 1\}$ represents whether the TF truly binds the DNA or not, based on the ChIP-seq experiment result; and σ_i is the output of the sigmoid function from the last layer. The CNN was trained using Adam optimizer [31] with batch size 1000, and the training was stopped when the validation loss did not decrease for over 100 epochs.

2.3.4 CNN learned motif extraction using a simulated annealing method

We used simulated annealing technique [69, 70], a Markov Chain Monte Carlo sampling method, to perform probabilistic optimization of the CNN-learned motif over the set of input sequences. We extracted and visualized the learned motif both for the CNN model (Section 4.3) and the tensor net model (Section 6.3) of TF SP1.

The simulated annealing algorithm samples sequences through a discrete-time inhomogeneous Markov chain with transition probabilities determined by a cost function J and a temperature parameter T. Specifically, sequences with lower J values are sampled with higher probability, and as the temperature decreases, the ratio of the sampling probability for sequences with lower J values to the sampling probability for sequences with higher J values increases.

To sample sequence $\boldsymbol{x}$ that maximizes the input of the sigmoid function in the last layer of the trained CNN, we used the cost function $J(\boldsymbol{x}) = -y(\boldsymbol{x})$, where $y(\boldsymbol{x})$ denotes the pre-activation of the output neuron. For the trained CNN, we initialized 50 instances of simulated annealing at the 50 elements of the test set (in the CNN model of SP1, the test set was SP1 A549 chr1 ChIP-seq dataset) predicted to have the highest pre-activation. The pseudocode for the simulated algorithm is given as Algorithm 1; $\boldsymbol{x}_{(n-1)}$ is the input sequence at the n^{th} iteration, d is the initial temperature, and $N_{iter} - 2$, N_{sample}, $N_{interval}$ are the iteration total number, sample size and interval size, respectively. At each iteration, we only changed one nucleotide in the sequence, while the quantile-normalized DNase-seq signal was held unchanged.

The initial temperature d is chosen by empirical observation of the distribution of the cost function J, in particular, by considering the height of communication of local minima [69]. Although we could not guarantee obtaining global minima at the end of the simulation, we still wanted to make sure that the

Algorithm 1 Simulated Annealing

Given: $\boldsymbol{x}_0$, $J(\boldsymbol{x})$, d, N_{iter}, N_{sample}, $N_{interval}$
for n in $(1, 2, \ldots, N_{iter} - 2)$ **do**
  $T = d/\log(n+1)$
  $i \sim Unif(\{0, \ldots, 1001\})$ ▷ select one base index in the input sequence
  $\boldsymbol{x}_{proposed} \equiv \boldsymbol{x}_{n-1}$
  $(\boldsymbol{x}_{proposed})_i \sim Unif(\{\text{A,C,G,T}\} - \{(\boldsymbol{x}_{n-1})_i\})$ ▷ replace the nucleotide with a randomly sampled one
  $u \sim Unif([0, 1])$
  if $exp(-\frac{(J(\boldsymbol{x}_{proposed}) - J(\boldsymbol{x}_{n-1})}{T}) > u$ **then**
    $\boldsymbol{x}_n = \boldsymbol{x}_{proposed}$ ▷ Accept
  else
    $\boldsymbol{x}_n = \boldsymbol{x}_{n-1}$ ▷ Reject
  end if
  return: $\{\boldsymbol{x}_n : (n > N_{iter} - (N_{sample})(N_{interval}) - 2)$ and $((n+1) \bmod N_{interval} = 0)\}$
end for

Markov chains could transition from shallow to deeper basins of the cost function J. We therefore selected the initial temperature d by estimating the minimum communication height for sequences near the local minima of the cost function. First, we supplemented each of the 50 test set inputs (A549 chr1) predicted to have the highest pre-activation $y(\boldsymbol{x})$ with 1000 inputs sampled uniformly and without replacement from the union of training, validation, and test sets. To remove the variation in $J(\boldsymbol{x})$ caused by DNase-seq, we fixed the DNase-seq signal of all 1000 samples to the corresponding maximum pre-activation test input DNase-seq signal. Next, for each of the test inputs and the supplemented 1000 samples, we calculated the difference $(J_{m,k} - \min_j(J_{m,j}))$, where $1 \leq k \leq 1001$, $1 \leq m \leq 50$, and $J_{m,k}$ denotes the k^{th} value of J among the 1001 samples associated with the test set input m. We then combined the calculated difference values, and ranked these 50000 difference values. We observed a transition from a rapid increase to a moderate increase for the 50000 ranked difference values. We thus estimated the fraction of inputs near strong local minima and chose the 1^{st} percentile of the ranked $U_{m=1}^{50}\{(J_{m,k} - \min_j(J_{m,j})) : 1 \leq k \leq 1001\}$ as a threshold, yielding the threshold d as the initial temperature. The values of the other parameters were $N_{iter} = 5 \times 10^5$, $N_{sample} = 10^4$ and $N_{interval} = 10$. Using these parameters, we performed simulated annealing starting from each of the 50 test inputs which were predicted to have the highest pre-activation values. For each of the 50 simulated annealing experiments, we monitored the minimum of $J(\boldsymbol{x})$ across the previous iterations versus the iteration number n. After the minimum of $J(\boldsymbol{x})$ stabilized, we recorded the sampled sequences $\boldsymbol{x}_n$ every 10 iterations starting from the 399999^{th} iteration for each simulated annealing experiment. For the CNN model of SP1 in Section 4.3, we chose the experiment with the lowest stable $\min(J(\boldsymbol{x}))$ out of the 50 experiments as our best scenario, and visualized the CNN-learned motif using the recorded sequences through WebLogo [79] 3.

2.4 Experimental validation

Up till now, we have identified candidate causal SNPs (Section 2.1), target genes (Section 2.2) and transcription factors (Section 2.3). For the best candidate (causal SNP, target gene, TF) triplets, we could then perform *in vitro* and *in vivo* experiments to validate the computational analysis.

Electrophoretic mobility shift assay (EMSA)

For the GWAS locus harboring rs648044, we performed an electrophoretic mobility shift assay (EMSA) experiment to confirm that the transcription factor MAFF preferentially binds rs648044-A allele. EMSA is also known as gel shift assay, and is a common technique to determine if a protein or a mixture of proteins could bind to a DNA or RNA sequence. In our experiment, we took positive control sequence, negative control sequence, 81bp flanking sequence containing rs648044 risk allele A, 81bp flanking sequence containing rs648044 non-risk allele G as inputs, and revealed the binding of MafF on positive control sequence and the flanking sequence containing rs648044 risk allele A (Section 3.3, Appendix B.2 and B.3).

RNA interference (RNAi) experiment

We also performed an *in vivo* RNA interference (RNAi) experiment to assess the effect of MAFF mRNA knockdown on target gene expression. RNAi is a gene-silencing process with exogenous or endogenous double-stranded RNA (dsRNA) involved. For RNAi initiated by exogenous dsRNA, the ribonuclease protein Dicer [80] binds and cleaves short hairpin RNAs (shRNAs), resulting in short double-stranded fragments [81, 82, 83, 84] - small interfering RNAs (siRNAs), which are then divided into single strand RNAs. The guide strand is then integrated into an RNA-induced silencing complex (RISC). For post-transcriptional gene silencing, the guide strand RNA in RISC pairs with a complementary sequence in the targeted mRNA, which is subsequently cleaved by a catalyst in RISC [85]. In the experiment of our study, shRNA was transferred into cells of a human *IDH1*$^{\mathrm{R132H}}$ mutant, *TERT* promoter-mutant, 1p/19q-codeleted oligodendroglioma cell line heterozygous at rs648044, and the expression level of putative target genes *ZBTB16* and *NCAM1* upon MAFF knockdown was measured (Section 3.3, Appendix B.4).

Proposed experiment - Allele specific chromatin immunoprecipitation coupled with quantitative PCR (ChIP-qPCR)

Besides EMSA and RNA interference experiments, here we also propose allele specific chromatin immunoprecipitation coupled with quantitative PCR (ChIP-qPCR) experiment to validate the allele specific binding of the candidate TFs [86, 87, 88]. ChIP-qPCR is a technique that could be used to indicate the interaction

of transcription factors and genomic binding sites [89]. The DNA sequences in ChIP-qPCR experiment rely on the primers to amplify. Allele specific ChIP-qPCR could further reveal the different enrichment level of two alleles of a SNP by designing the "allele-specific mismatch amplification mutation assays primers" [86, 90]. We thus could design two primers for two DNA sequences containing the risk and non-risk allele, and quantitatively measure the enrichment level of two alleles. We could further perform high-throughput sequencing to assess the allele-specific read coverage of the binding sites, which could help to validate the allele binding preference of TFs [87].

Chapter 3

The functional role of 11q23.2 variant rs648044 - *ZBTB16* locus

The work done in this chapter has been published in Manjunath†, Yan† [1] *et al.*, Neuro-oncology, 2020 [51]. The computational part of this chapter was performed by the author in collaboration with Dr. Mohith Manjunath, who performed the ATAC-seq allele specific read counts analysis (in Section 3.2, Appendix B.1), *CIC* mutation analysis (in Section 3.3, Appendix B.6). Dr. Mohith Manjunath also contributed to PLAC-seq analysis (in Section 3.1), MAFF motif permutation analysis (in Section 3.2), and proposed the potential interaction of ZBTB16 with *CIC* through analyzing the ChIP-seq data of ZBTB16 (Section 3.4). The EMSA experiment was performed by Yeoan Youn in Professor Paul R. Selvin's lab at UIUC (in Section 3.3, Appendix B.2), and the RNAi experiment was performed by Dr. Kristen L. Drucker in Professor Robert B. Jenkins' lab at Mayo Clinic (in Section 3.3, Appendix B.4, B.5). The positive and negative control sequences in the experiments were provided by the author (Appendix B.3).

3.1 The lead SNP rs648044 modulates the expression of *ZBTB16* through chromatin looping

The lead GWAS SNP rs648044 contained no other SNP in high LD within its haplotype block (Figure 3.1) and was thus our candidate causal variant. As functional variants often interact with their target genes through active regulatory elements, we examined the epigenetic landscape surrounding the SNP in brain-related tissues and cell lines. Independent ATAC-seq [54, 55] and DNase-seq datasets confirmed the SNP to be located within an open chromatin region in TCGA LGG samples, oligodendrocytes and fetal brain tissue samples (Figure 3.2A). Histone H3 lysine 4 mono-methylation (H3K4me1) and H3K27 acetylation (H3K27ac) ChIP-seq in fetal brain and dorsolateral prefrontal cortex tissues showed an active enhancer activity at the location (Figure 3.2A), as also annotated by REMC (Figure 3.3A).

We further identified the target gene as described below. The eQTL analysis using the genotypes and gene expression data of TCGA LGG samples suggested *NCAM1* and *ZBTB16* to be the top candidate target

[1] †co-first authors, contributed equally.

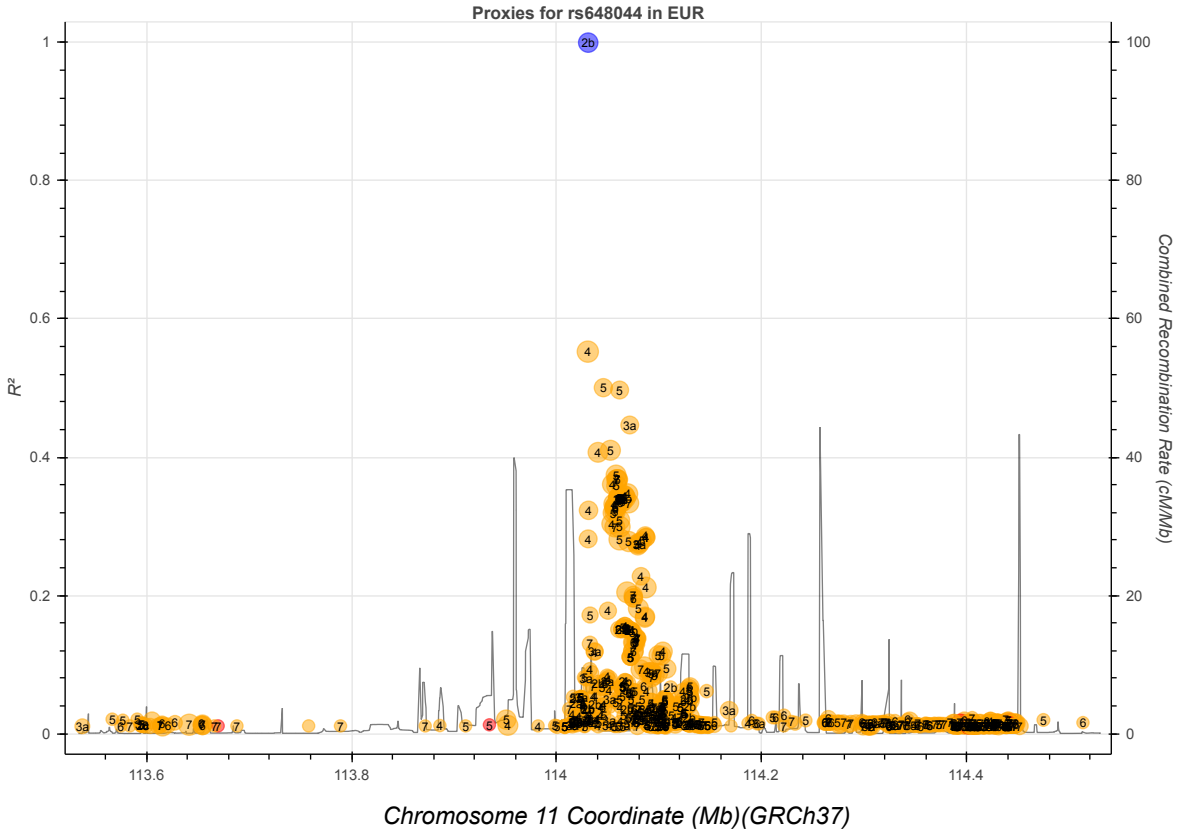


Figure 3.1: The lead GWAS SNP rs648044 contained no other SNP in high LD within its haplotype block. The figure was downloaded from LDlink [58] with rs648044 as the query variant and European (EUR) as the query population. R^2 of each SNP and the combined recombination rate are shown in the figure. Purple, yellow and red circles denote query variant, non-coding variants and coding variants, respectively. The numbers (ranging from 1 to 7, from RegulomeDB [91]) in the circles represent the regulatory potential of the variants, with 1 denoting the highest and 7 the lowest; the variant's regulatory potential from RegulomeDB is unknown if there is no number in the circle.

genes (Figure 3.2B, Figure 3.3B, *NCAM1* $P = 0.0054$ in the combined "IDH^{mut} only" and triple-positive group). *NCAM1* is located about 1.1 Mb away from *ZBTB16*. H3K4me3 PLAC-seq confirmed a physical looping interaction only between the active *ZBTB16* promoter and the enhancer harboring rs648044 in oligodendrocytes [55] (Figure 3.4A). We thus prioritized *ZBTB16* for further analysis. Correlation analysis between *ZBTB16* normalized expression values and genotype status at rs648044 in different molecular groups found a significant association in the combined group of "IDH^{mut} only" and triple-positive ($P = 0.0118$, FDR = 0.124; Figure 3.3B). The expression level of *ZBTB16* was suppressed by the rs648044-A risk allele, indicating that *ZBTB16* might act as a tumor suppressor. Consistent with this hypothesis, *ZBTB16* encodes a zinc-finger TF [92] implicated in inhibiting proliferation, metastasis or epithelial-mesenchymal transition in multiple cancers and is genetically lost in metastatic castration-resistant prostate cancer [93], supporting

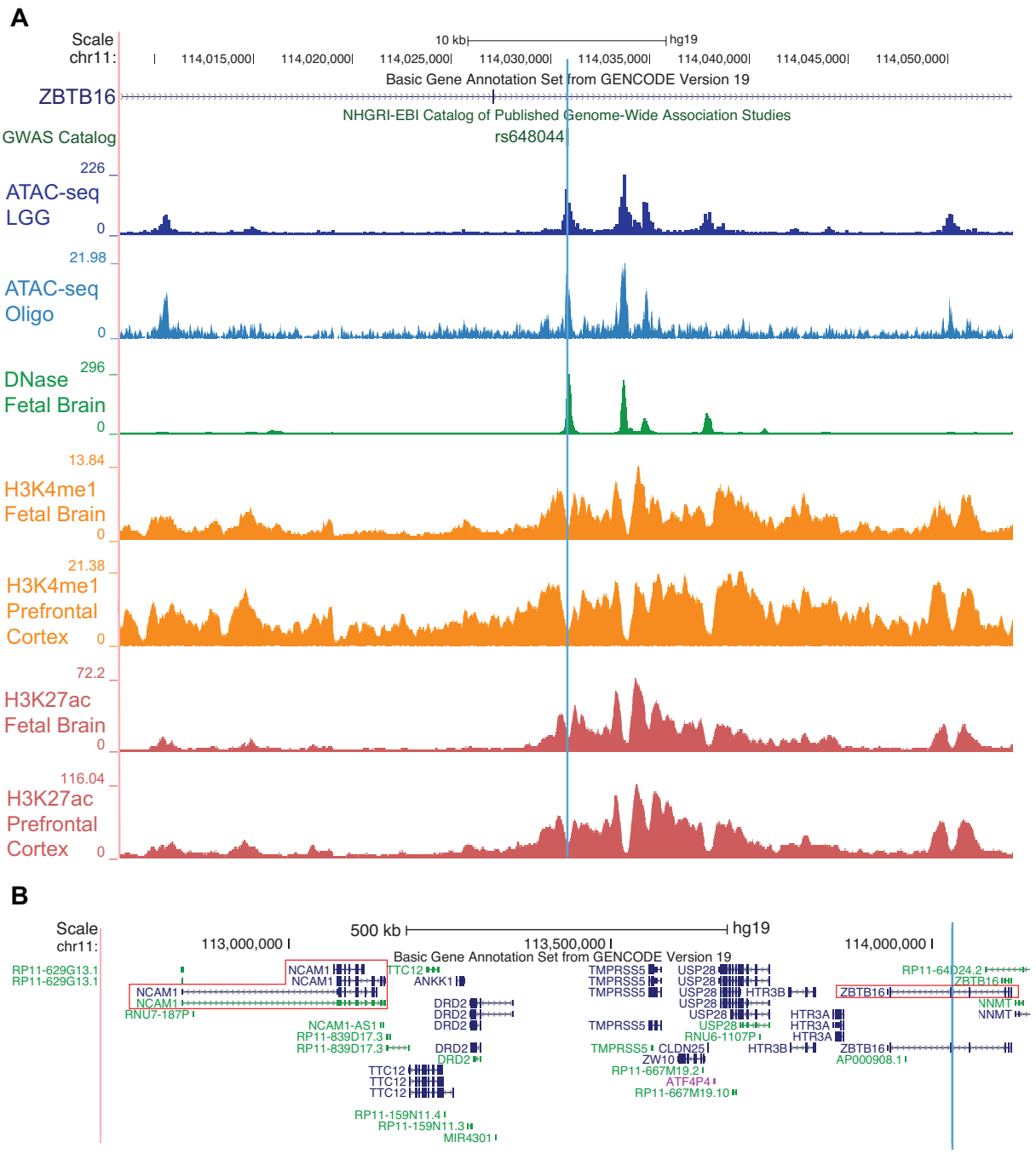


Figure 3.2: *ZBTB16* intronic enhancer harboring rs648044 modulates mRNA expression of nearby genes. (A) A snapshot of the *ZBTB16* locus where the GWAS SNP rs648044 is denoted by a blue vertical line. The shown epigenomic tracks are: TCGA-LGG ATAC-seq [54], oligodendrocytes ATAC-seq [55], and REMC data in fetal brain and prefrontal cortex. (B) A zoomed-out view of the *ZBTB16* locus encompassing the eQTL target genes, *NCAM1* and *ZBTB16*, indicated by red boxes.

its tumor suppressor role [94, 95, 96].

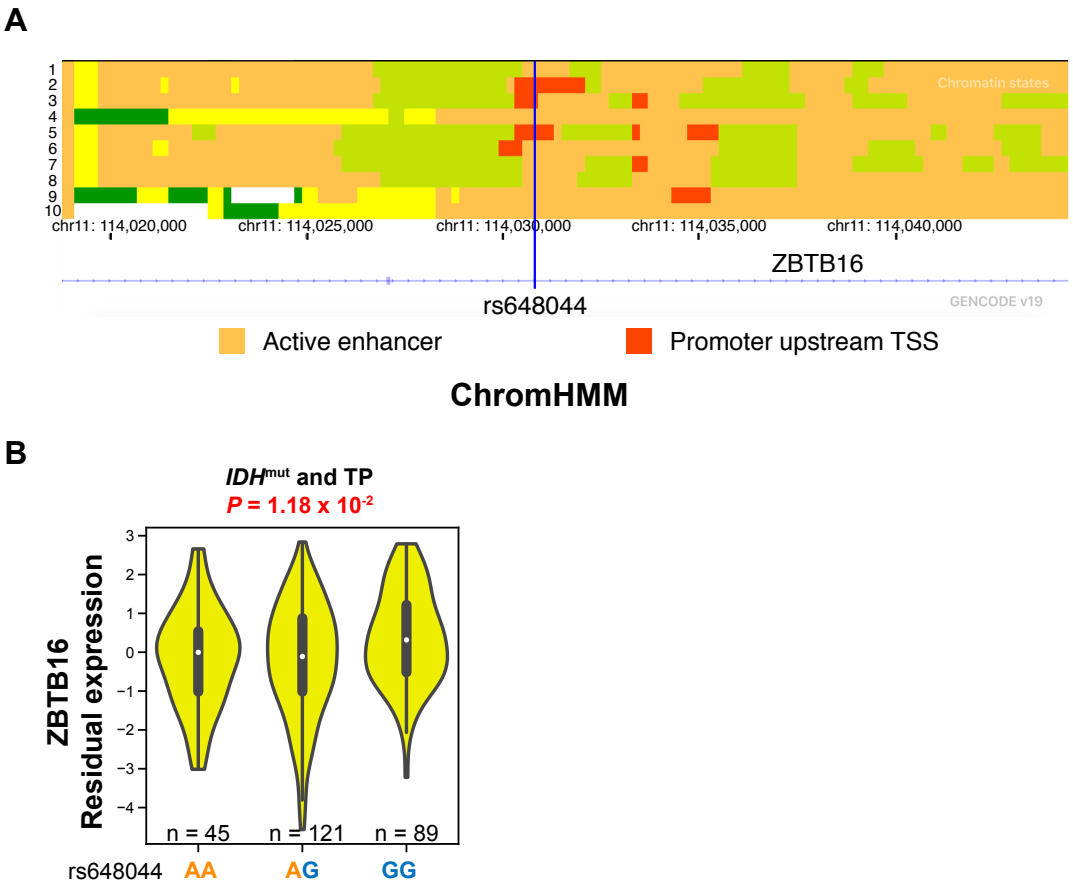


Figure 3.3: The GWAS SNP rs648044, located in an intron of *ZBTB16*, modulates *ZBTB16* mRNA expression. (A) ChromHMM [53] tracks of 10 brain tissue samples from Roadmap Epigenomics Mapping Consortium (REMC) database [28, 29] for the region harboring the GWAS SNP rs648044. The samples are Brain Angular Gyrus, Anterior Caudate, Cingulate Gyrus, Germinal Matrix, Hippocampus Middle, Inferior Temporal Lobe, Dorsolateral Prefrontal Cortex, Substantia Nigra, Fetal Brain Female and Fetal Brain Male. (B) eQTL result for rs648044 and *ZBTB16* in the combined TCGA-LGG "IDH^{mut} only" and triple-positive (TP) group. Throughout the text, the risk and non-risk alleles of a SNP are colored orange and blue, respectively.

3.2 rs648044 likely perturbs the binding affinity of MAFF

We next sought to identify the TF whose binding affinity might be perturbed by rs648044. We first utilized known TF binding motifs to perform *in-silico* TF binding affinity perturbation analysis based on a sequence permutation test (Section 2.3.1, Appendix A.2). For each candidate TF, we then computed molecular group-wise Pearson correlation coefficient between the TF and *ZBTB16* expression levels stratified into three genotype groups of rs648044 (Section 2.3.2). Based on the eQTL finding that *ZBTB16* expression was lower in the risk group (AA genotype), we expected a candidate repressor TF to have higher binding affinity towards the risk allele A and show a greater negative correlation with *ZBTB16* in the risk group compared

to the GG genotype group; conversely, we expected a candidate activator TF to have lower binding affinity towards the risk allele and show a weaker positive correlation with *ZBTB16* in the risk group. ATAC-seq data in TCGA LGG samples showed a significant skew toward the rs648044-A risk allele, indicating that the TF might act as a repressor ($P = 0.010$, Fisher's method for combining binomial test P-values; Table B.1; Section 2.3.3, Appendix B.1). These criteria together identified MAFF as the top candidate TF for further experimental validation. MAFF is a member of the small MAF basic leucine zipper TFs that can homodimerize and repress target genes. Its motif [97] clearly preferred the risk allele A (Figure 3.4B; permutation test $P = 0.0029$, Section 2.3.1, Appendix A.2), and the structure of expression correlation showed attenuation of the negative correlation between *MAFF* and *ZBTB16* in the AG and GG genotypes that were predicted to weaken the affinity of MAFF to DNA (Figure 3.4C).

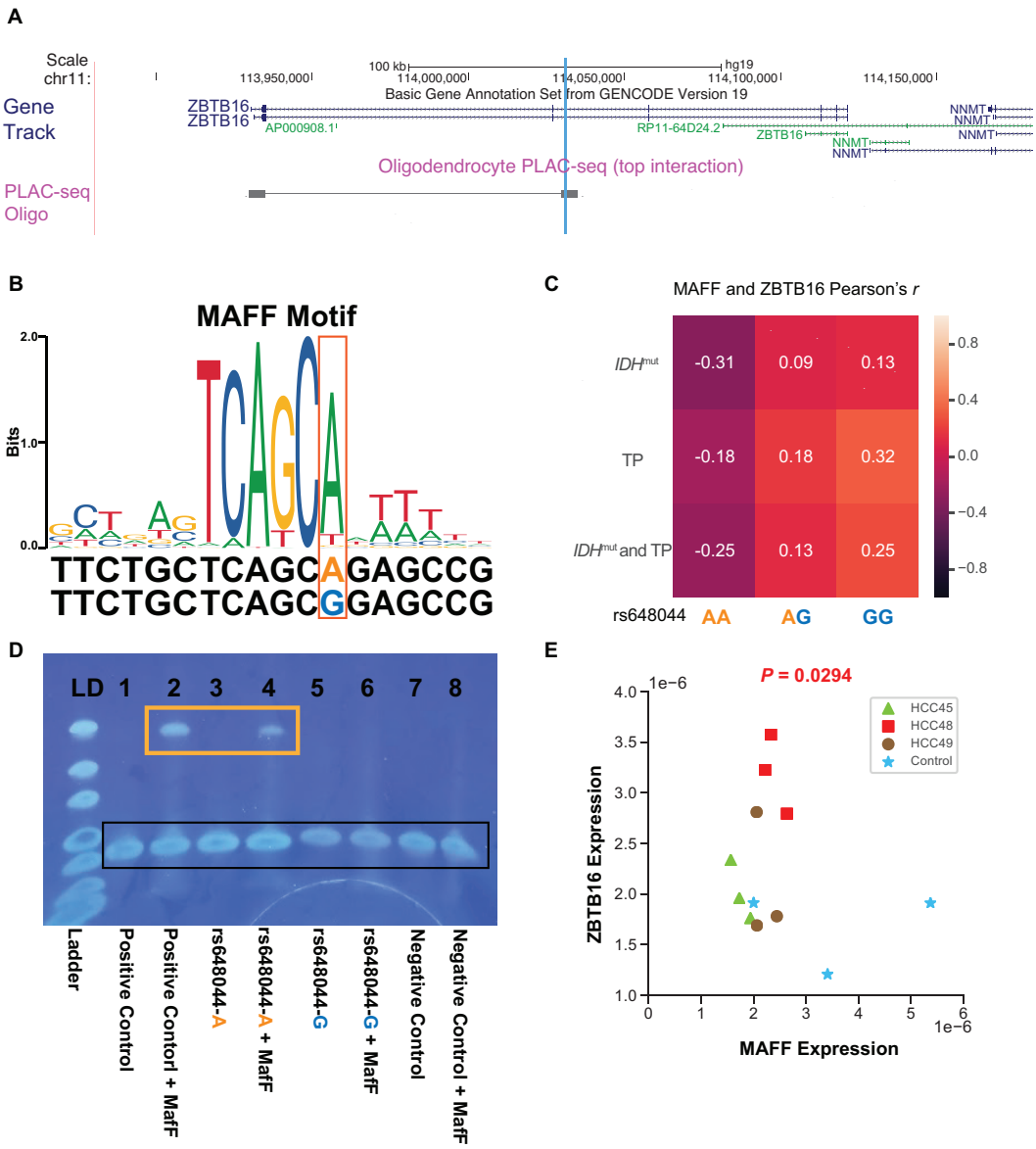


Figure 3.4: The GWAS SNP rs648044 likely perturbs the binding affinity of MAFF that represses *ZBTB16*. (A) Oligodendrocyte PLAC-seq track [55] showing the rs648044 locus (blue vertical line) interacting with the *ZBTB16* promoter, about 100 kb away. (B) JASPAR [97] motif logo of the predicted TF MAFF and two variants of the flanking sequence harboring rs648044-A and rs648044-G alleles. Throughout the text, the risk and non-risk alleles of a SNP are colored orange and blue, respectively. (C) Pearson's correlation coefficient between *ZBTB16* and *MAFF* in the combined "*IDH*$^{\text{mut}}$ only" and TP group, "*IDH*$^{\text{mut}}$ only" subgroup and TP subgroup. (D) Gel picture from the EMSA experiment showing a ladder ("LD") and eight lanes using the mixture of the recombinant MafF protein and 4 different DNA sequences (Appendix B.2, Appendix B.3): 81bp positive control ("PC") sequence, 81bp sequence flanking rs648044-A, 81bp sequence flanking rs648044-G and negative control ("NC") sequence. The lower molecular weight bands in black box correspond to free DNA. Orange box highlights the bands of MafF-bound DNA, corresponding to the results of "positive control DNA + MafF" and "rs648044-A flanking sequence + MafF". (E) MAFF RNAi knockdown experiment results, showing a significant increase in *ZBTB16* mRNA expression after MAFF knockdown. One-sided t-test P-value between the control group and the combined group of three independent shRNA clones is shown on the top.

3.3 RNA interference and EMSA experiments

To confirm that MAFF preferentially binds the rs648044-A allele, we performed an electrophoretic mobility shift assay (EMSA) (Figure 3.4D, Appendix B.2). We detected binding of MAFF on positive control DNA (from a top consensus ChIP-seq peak region in HepG2, K562 and HelaS3, Appendix B.3; lane 2) and the sequence containing the risk A allele (Appendix B.2; lane 4), but not on the sequence containing the alternative G allele (Appendix B.2; lane 6) and negative control DNA (a permuted sequence with no MAFF core binding motif, Appendix B.3; lane 8). Knockdown of MAFF using shRNA in a cell line – derived from an *IDH1*$^{\mathrm{R132H}}$ mutant, *TERT* promoter-mutant, 1p/19q-codeleted (triple positive) oligodendroglioma patient and heterozygous at rs648044 (Appendix B.4) – led to a significant increase in *ZBTB16* mRNA expression compared to non-target controls (Figure 3.4E; $P = 0.0294$, two-group one-sided t-test), but not in *NCAM1* mRNA expression (Figure B.1; $P = 0.37$, two-group one-sided t-test). These results support our prediction that MAFF preferentially binds the risk allele rs648044-A and represses the putative tumor suppressor *ZBTB16*. We further analyzed the prevalence of *CIC* mutations in the context of rs648044 genotypes (Appendix B.6), as CIC is an important tumor suppressor frequently mutated in *IDH*$^{\mathrm{mut}}$ gliomas. *CIC* inactivating mutations tended to occur more frequently in the homozygous non-risk GG genotype than the combined AA and AG genotypes in TCGA triple-positive gliomas (Odds ratio 2.0, Fisher's exact test $P = 0.076$; Table 3.1), although statistical significance could not be reached, potentially due to small sample size. This finding suggested that the predicted suppression of *ZBTB16* by the risk rs648044-A allele could be an alternate mechanism for LGG tumorigenesis in *CIC* wild-type gliomas.

CIC status	GG	AA+AG
Mutant	16	25
Wildtype	15	48

Table 3.1: Inactivating mutation status of *CIC* in the triple-positive group stratified into non-risk (GG) and risk (AA+AG) genotypes of rs648044.

3.4 ZBTB16 and *CIC*

We have shown that the 11q23.2 GWAS SNP rs648044 may modulate the expression of *ZBTB16* by perturbing the binding affinity of MAFF. Although ENCODE ChIP-seq data show a MAFF peak (q-value $= 3.1 \times 10^{-4}$) covering the SNP rs648044 in K562 cells, as well as a similar MAFK peak (q-value $= 1.6 \times 10^{-4}$) in HepG2 cells, further studies are needed to confirm the allele-specific binding of MAFF at rs648044 in glioma cells, as predicted by our computational analysis and *in vitro* data. ZBTB16 has been shown to regulate

self-renewal and differentiation of hematopoietic stem cells, mainly acting as a transcriptional activator and antagonized by a noncanonical function of the histone methyltransferase EZH2 [98]. It also acts as a tumor suppressor in prostate cancer, melanoma, gallbladder cancer and leukemia [95, 96, 99, 100]. Although no ZBTB16 ChIP-seq data are currently available in oligodendrocytes, ChIP-seq data in human mesenchymal stem cells [101], endometrial stromal cells [102] and acute myelogenous leukemia cells [98] show ZBTB16 binding the *CIC* promoter in these cell types (Figure 3.5, Appendix B.7). The mRNA expression level of *ZBTB16* is also highly correlated with that of *CIC* in prefrontal cortex (Spearman's $\rho = 0.65$, GTEx v8), supporting that *CIC* is likely a direct transcriptional target of ZBTB16. Importantly, *CIC* is one of the most commonly mutated genes in IDH^{mut} oligodendrogliomas and located on chromosome 19q, which is often codeleted with chromosome 1p in oligodendrogliomas. These observations thus suggest a potentially important interaction network involving the regulation of *CIC* by ZBTB16 and disruption of this interaction by rs648044 in the tumorigenesis of LGG. The fact that *CIC* mutation shows a trend of being more frequent in the homozygous non-risk GG genotype of rs648044, where the expression level of ZBTB16 is elevated, is consistent with this potential interaction between the two tumor suppressors. However, the sample size of patients in our study may be too small to understand the genetic interactions accurately; furthermore, some patients having the non-risk GG genotype of rs648044 may have mutations in other genes or harbor other risk SNPs, leading to alternate mechanisms of LGG pathogenesis [12, 59].

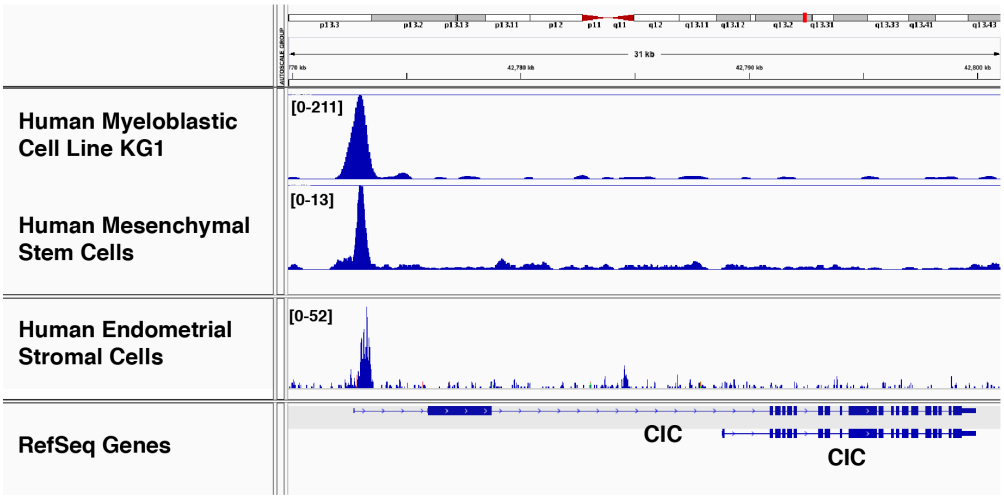


Figure 3.5: ZBTB16 binds the *CIC* promoter in multiple cell types. A snapshot at the *CIC* locus from the Integrative Genomics Viewer [103] showing ZBTB16 ChIP-seq data in human acute myelogenous leukemia cell line KG1 [98], mesenchymal stem cells [101] and endometrial stromal cells [102].

3.5 Conclusion

We have thus presented the (MAFF, rs648044, *ZBTB16*) triplet that might contribute to the pathogenesis of LGG. Through concrete computational analysis we proposed that MAFF preferentially binds rs648044 risk allele A and represses the expression of the putative tumor suppressor *ZBTB16* (Figure 3.6). We performed *in vitro* EMSA experiment which showed the binding preference of MAFF to rs648044-A allele. We also performed *in vivo* RNAi experiment, and validated the increase of *ZBTB16* expression upon MAFF knockdown. We then analyzed the prevalence of *CIC* mutations with regard to rs648044 genotype, and found that *CIC* showed a trend of being more frequently mutated in the non-risk rs648044 GG genotype group where *ZBTB16* expression level was elevated. Since ZBTB16 binds the promoter of *CIC* in three cell types and *CIC* is an important tumor suppressor frequently mutated in oligodendrogliomas, the predicted (MAFF, rs648044, *ZBTB16*) triplet might be involved in a larger interaction network contributing to LGG tumorigenesis.

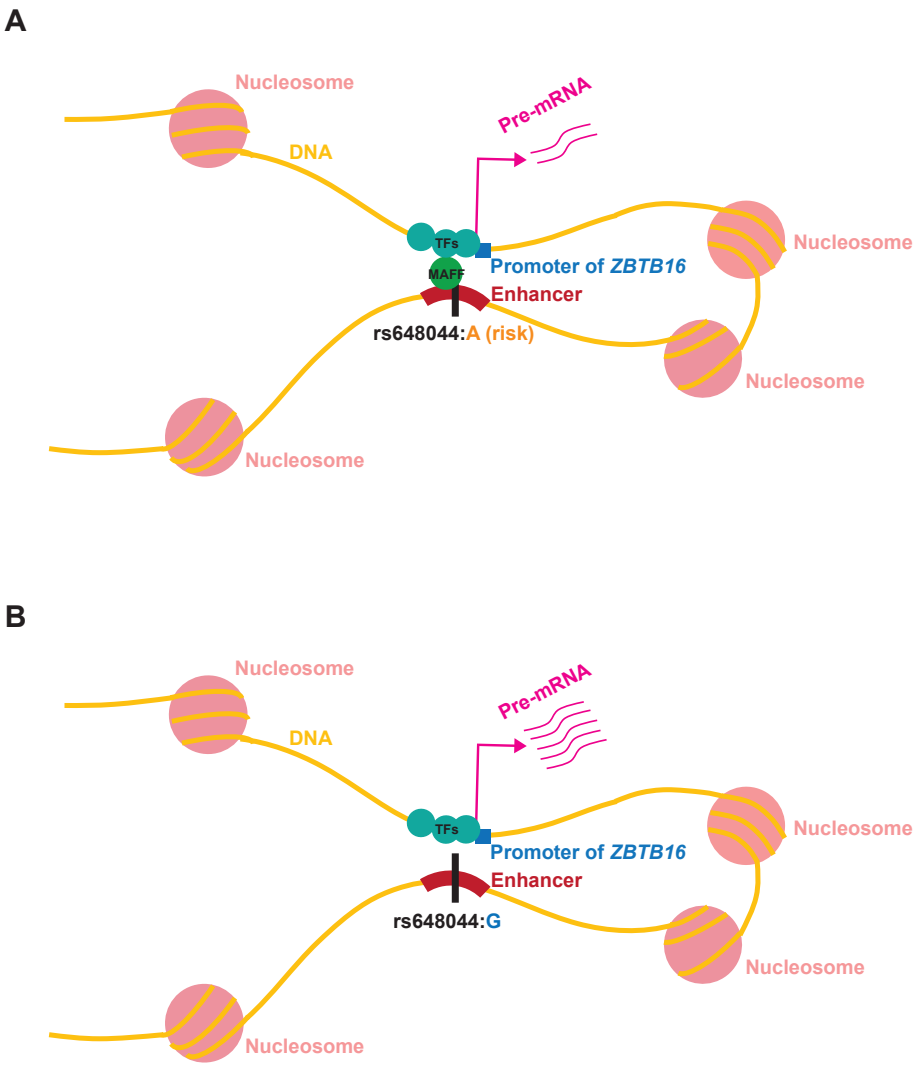


Figure 3.6: Illustrative plot showing the GWAS SNP rs648044 modulates the expression of *ZBTB16* through chromatin looping and perturbing the binding affinity of MAFF. (A) MAFF preferentially binds rs648044 risk allele A, and acts as a transcription repressor to repress the expression of *ZBTB16*. (B) rs648044 non-risk allele G weakens the binding affinity of MAFF, and as a result the expression level of *ZBTB16* is elevated.

Chapter 4

The functional analysis of 11q23.3 variant rs12803321 - *PHLDB1* locus

In this chapter, we present the functional analysis of 11q23.3 GWAS variant rs12803321. The work in this chapter was mainly done by the author, and has been published in Manjunath†, Yan† [1] *et al.*, Neuro-oncology, 2020 [51].

4.1 eQTL and phased ASE analyses implicate *PHLDB1* as a candidate target gene

We next applied our computational framework to the locus containing rs12803321 (reference allele: G (risk), alternative allele: C), one of the most significant LGG GWAS SNPs. The SNP rs12803321, located in the first intron of Pleckstrin Homology Like Domain Family B Member 1 (*PHLDB1*) (Figure 4.1C), was reported to be significantly associated with the "IDH^{mut} only" subgroup [104, 105]. An eQTL analysis of 71 genes within 4Mb of rs12803321 in "IDH^{mut} only" subgroup (Section 2.2.3) identified *PHLDB1* and Trehalase (*TREH*) as the top candidate target genes (*PHLDB1* $P = 2.5 \times 10^{-9}$, FDR $= 1.82 \times 10^{-7}$; *TREH* $P = 8 \times 10^{-5}$, FDR $= 2.84 \times 10^{-3}$; Figure 4.1A,B). The number of risk alleles was anticorrelated with the expression level of *PHLDB1* and *TREH* adjusted for covariates (Section 2.2.3). Since *TREH* expression was low (zero RSEM in 68 patients out of total 193), we prioritized *PHLDB1* for further analysis. We analyzed the allele-specific transcription pattern of *PHLDB1* using TCGA RNA-Seq raw reads and the exonic SNPs' phased haplotype information (Section 2.2.4). There were 20 exonic SNPs with more than 5 cases in the "IDH^{mut} only" group having a heterozygous genotype at both rs12803321 and the exonic SNP. Wilcoxon signed-rank sum test on the RNA-Seq read counts from the two chromosomes [68] detected a statistically significant skew at 9 exonic SNPs out of 20 ($P < 0.05$). All these 9 SNPs showed higher transcription emanating from the rs12803321-C haplotype (Figure C.1). These results together demonstrated that the risk allele rs12803321-G was associated with decreased expression of *PHLDB1* in the "IDH^{mut} only" group.

[1] †co-first authors, contributed equally.

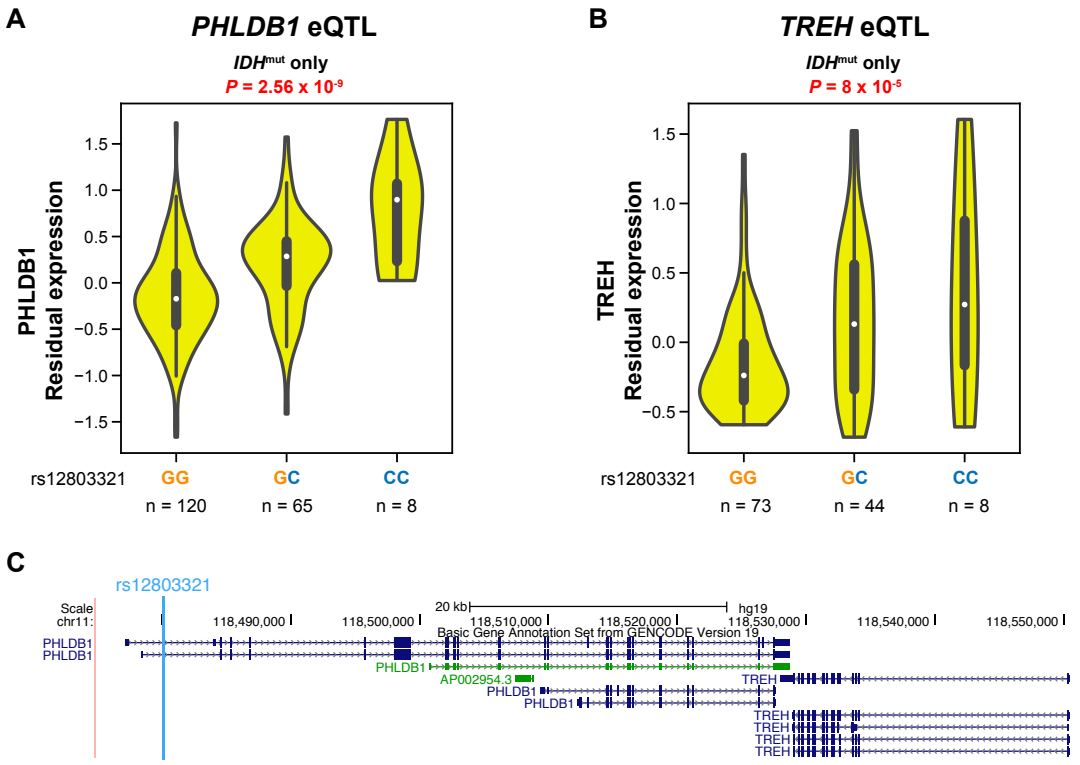


Figure 4.1: eQTL analysis indicates *PHLDB1* and *TREH* as the top candidate target genes. (A) eQTL result for rs12803321 and *PHLDB1* in the TCGA-LGG "IDH^{mut} only" subgroup. (B) eQTL result for rs12803321 and *TREH* in the TCGA-LGG "IDH^{mut} only" subgroup. (C) Gene track showing the GWAS SNP rs12803321 and candidate target genes, *PHLDB1* and *TREH*. rs12803321 is denoted by a blue vertical line.

4.2 Candidate causal SNP rs12225399 perturbs the binding affinity of SP1/SP2

We next prioritized candidate functional SNPs using epigenomic data. There were three SNPs in high LD with rs12803321 (Table C.1, Section 2.1.2): rs67307131 ($r^2 = 0.98$), rs12225399 ($r^2 = 0.97$) and rs7125115 ($r^2 = 0.90$). The GWAS SNP and all three high LD SNPs were located in open chromatin and active enhancer regions, as assessed by the fetal brain DNase-seq, TCGA LGG ATAC-seq [54], oligodendrocyte ATAC-seq [55], and prefrontal cortex histone modification (H3K4me1, H3K27ac) ChIP-seq data (Figure 4.2A). Motif analysis using FIMO [71] yielded candidate TFs whose binding affinity might be perturbed by any of the above four SNPs (Appendix C.3, C.4, C.5, C.6). Further filtering the TF list through TF-target gene expression correlation analysis (Section 2.3.2), we determined rs12225399 to be the best candidate causal SNP, and SP1/SP2 the top candidate TFs: first, rs12225399 was located near a local peak center

in TCGA LGG and oligodendrocyte ATAC-seq (Figure 4.2A,B); second, sequence perturbation analyses demonstrated that the rs12225399-C allele, in phase with the rs12803321-C allele, created a high-scoring SP1/SP2 binding motif, whereas the rs12225399-G allele significantly perturbed the motif (FIMO SP1 $P = 4.25 \times 10^{-5}$, Figure 4.3A; FIMO SP2 $P = 5.53 \times 10^{-5}$, Figure 4.3B; permutation test SP1 $P = 0.015$, SP2 $P = 0.0023$; Section 2.3.1, Appendix A.2); third, Pearson correlation coefficient between SP2 and *PHLDB1* in "IDH^{mut} only" group was highest in the rs12225399-CC genotype ($r = 0.40$) and decreased in rs12225399-GC ($r = 0.26$) and rs12225399-GG genotypes ($r = 0.23$) (Figure 4.3C). The correlation between SP1 and *PHLDB1* did not show the same trend as SP2 and *PHLDB1* (Figure 4.3D); however, since SP1 and SP2 recognize similar sequences (Figure 4.3A,B), we could not rule out SP1 as not being functional at the SNP. The high LD SNP rs7125115 was not selected as a candidate causal SNP, because our analysis did not yield a good candidate TF (Appendix C.3). These results together implied that the rs12225399-C allele likely increased the binding affinity of SP1/SP2, functioning as transcription activators to enhance the expression of *PHLDB1*.

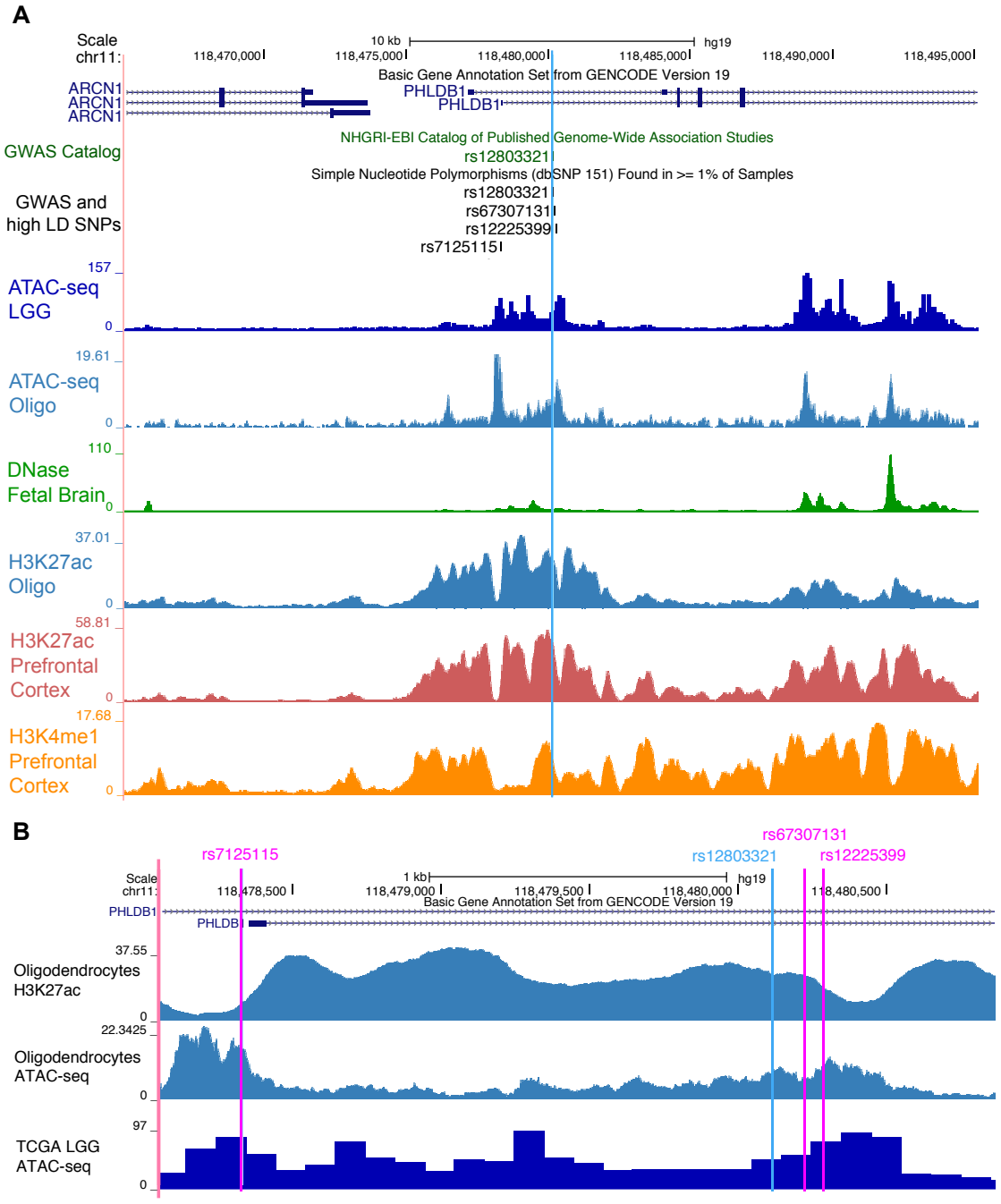


Figure 4.2: The epigenomic landscape of the region harboring the GWAS SNP rs12803321. (A) A snapshot of the *PHLDB1* locus where the GWAS SNP rs12803321 is denoted by a blue vertical line. The top three tracks are: basic gene annotation set from GENCODE version 19, LGG GWAS SNPs from GWAS catalog [106] and SNPs in high LD with rs12803321 from the Single Nucleotide Polymorphism Database [107] (dbSNP 151). The lower epigenomic tracks are: TCGA-LGG ATAC-seq [54], oligodendrocytes ATAC-seq [55], oligodendrocytes H3K27ac [55] and REMC data in fetal brain and prefrontal cortex. (B) Enlarged view of the epigenomic landscape of the region harboring the GWAS SNP rs12803321 and three high LD SNPs. Tracks from top to bottom are the oligodendrocytes H3K27ac [55], oligodendrocytes ATAC-seq [55] and TCGA-LGG ATAC-seq signals [54].

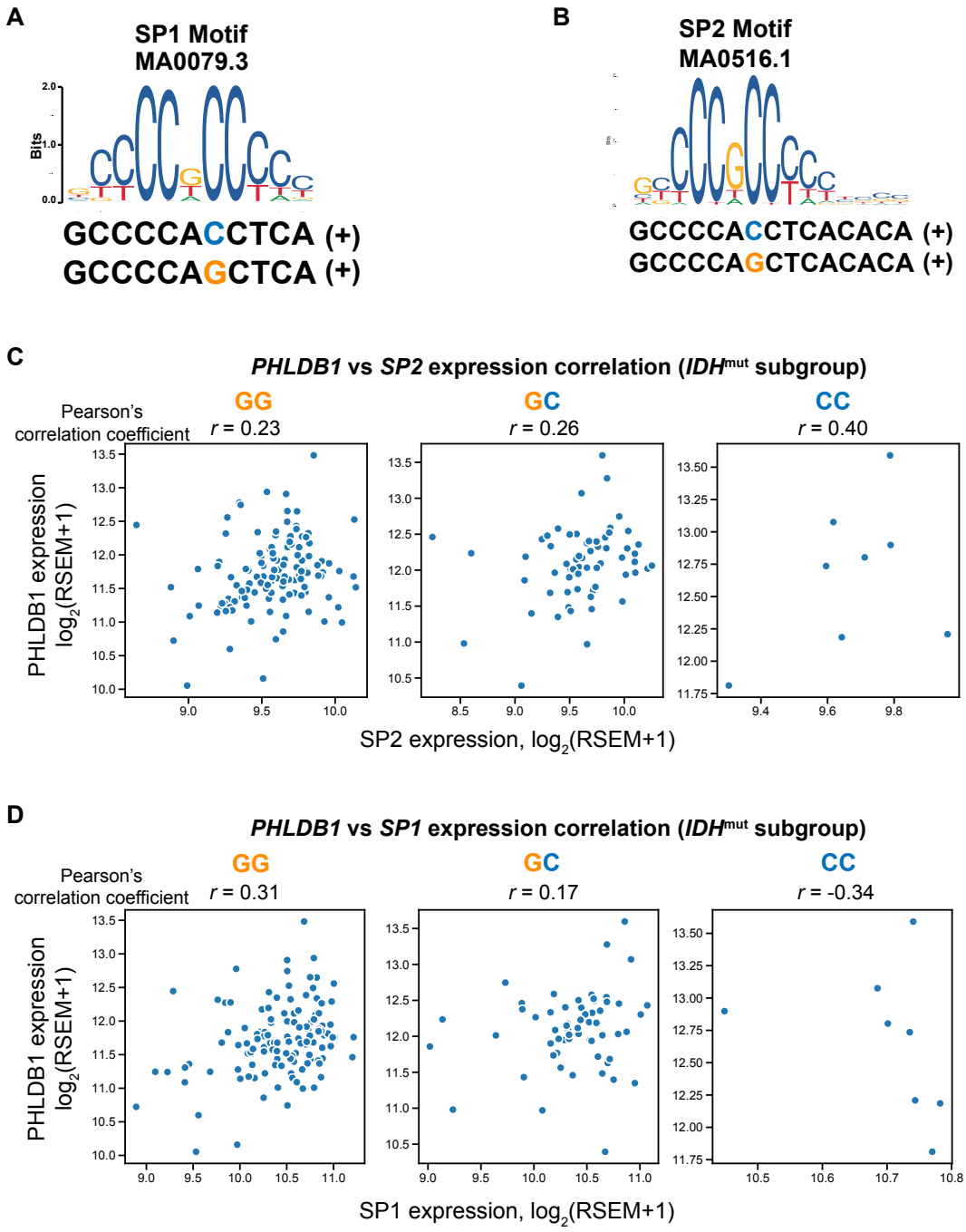


Figure 4.3: The high LD SNP rs12225399 likely modulates *PHLDB1* expression by perturbing the binding affinity of SP1/SP2. (A) SP1 motif logo MA0079.3 (JASPAR [24]) and two variants of the flanking sequence harboring rs12225399-C and rs12225399-G alleles. Allele C and G of rs12225399 are colored blue and orange, respectively. (B) SP2 motif logo MA0516.1 (JASPAR [24]) and two versions of the flanking sequence harboring the rs12225399-C and rs12225399-G alleles. (C) Scatter plots of *SP2* vs. *PHLDB1* expression in the three genotypes of rs12225399 in the TCGA-LGG "IDH^{mut}only" subgroup. Spearman's correlation coefficients between *SP2* and *PHLDB1* in the "IDH^{mut} only" group are: $\rho = 0.25$ (rs12225399-GG), $\rho = 0.26$ (rs12225399-GC), $\rho = 0.36$ (rs12225399-CC). (D) Scatter plots of *SP1* vs. *PHLDB1* expression in the three genotypes of rs12225399 in the TCGA-LGG "IDH^{mut} only" subgroup.

4.3 SP1 allele-specific binding prediction using CNN

Because of the lack of SP1/SP2 ChIP-seq data in brain cell types, we could not verify directly whether SP1/SP2 actually bound the predicted causal SNP. We thus applied a deep learning method to predict TF binding affinity in fetal brain samples (Section 2.3.3). Although SP2 was a better candidate, SP2 ChIP-seq data were available in only one ENCODE cell line, while SP1 ChIP-seq data were available in seven cell lines (H1-hESC, HEK293T, HepG2, Liver, K562, MCF-7, and A549; Table C.10). We thus trained a CNN for SP1 only, using sequence information and cell type-matched DNase-seq to predict the SP1 ChIP-seq signals.

4.3.1 Basic structure of constructed CNN and the training details

The structure of the constructed CNN is shown in Figure 2.5. The training details are stated below. We trained the CNN using SP1 ChIP-seq data from six cell lines/tissue (H1-hESC, HEK293T, HepG2, Liver, K562, MCF-7) and tested its performance using the cell line A549 (Table C.10). We obtained DNase-seq and ChIP-seq data in the above cell lines from ENCODE or REMC databases, and performed quantile normalization to reduce batch effect. We then collected all optimal ChIP-seq peak regions centered around peak centers to form the positive dataset (102934 samples) and selected an equal number of regions with no ChIP-seq peak to form the negative dataset. To increase the number of training samples and reduce overfitting, we then translated our initial positive and negative datasets by -20 bp, -10 bp, 0 bp, 10 bp, and 20 bp to form the translated positive and negative datasets. After removing the samples that fell into hg19 "blacklist" regions, we obtained a total of 514668 samples for the translated positive dataset and 514634 samples for the translated negative dataset. Thus, our CNN model has total parameter number 80141 (2440 parameters, convolutional layer; 76880 parameters, first dense layer; 810 parameters, second dense layer; 11 parameters, third dense layer), and total sample number 1029302. We then split the translated datasets into training and validation datasets with ratio 80% to 20% (training dataset: 823441 samples; validation dataset: 205861 samples). We tested the performance of the trained CNN using A549 chromosome 1 (chr1) positive and negative datasets (3785 samples for each) and calculated the receiver operating characteristic (ROC) area under curve (AUC) (Figure 4.4A). Finally, the trained CNN was used to predict the binding affinity of SP1 at rs12225399 using the REMC fetal brain DNase-seq samples as input (Table C.11).

4.3.2 Extraction of CNN-learned motif of SP1 using simulated annealing

We next extracted the CNN-learned motif of SP1 using the simulated annealing technique [69, 70]. Simulated annealing is a Markov Chain Monte Carlo sampling method, and was applied by Dr. Alex Finnegan for the

probabilistic optimization of CNN-predicted methylation rates over allowed input sequences [69]. Based on Dr. Alex Finnegan's method, we modified and used it to perform probabilistic optimization of the CNN-learned motif of SP1 over the set of input sequences. The details of our simulated annealing method is stated in Section 2.3.4. We summarize the basic steps and used parameters here.

We initiated 50 instances of simulated annealing at the 50 elements of A549 chr1 ChIP-seq test dataset predicted to have the highest pre-activation. We followed the pseudocode listed in Algorithm 1 with the following parameters: initial temperature $d = 2.24$; $N_{iter} = 5 \times 10^5$; sample size $N_{sample} = 10^4$ and interval size $N_{interval} = 10$. For each of the 50 simulated annealing experiments, we monitored the minimum of $J(\boldsymbol{x})$, which is defined as $J(\boldsymbol{x}) = -y(\boldsymbol{x})$, where $y(\boldsymbol{x})$ denotes the pre-activation of the output neuron, across the previous iterations versus the iteration number n. After the minimum of $J(\boldsymbol{x})$ stabilized, we recorded the sampled sequences every 10 iterations starting from the 399999^{th} iteration for each simulated annealing experiment. We then chose the experiment with the lowest stable $min(J(\boldsymbol{x}))$ out of the 50 experiments as our best scenario, and visualized the CNN-learned motif using the recorded sequences through WebLogo [79] 3.

The optimal motif extracted from the simulated annealing method closely resembled the known SP1 motif [24] (Figure 4.4C, Figure 4.3A). We thus could use the trained CNN model for allele-specific binding prediction at the two alleles of rs12225399 in the fetal brain.

4.3.3 The CNN model predicted differential binding of SP1

We first tested the performance of the trained CNN model using A549 chr1 positive and negative datasets (3785 samples for each), and calculated the receiver operating characteristic (ROC) area under curve, resulting in 0.95 (Figure 4.4A). Moreover, we confirmed that the optimal CNN-learned motif, extracted via a simulated annealing method, closely resembled the known SP1 motif [24] (Figure 4.4C, Figure 4.3A). The trained CNN was then used to evaluate the impact of rs12225399 on SP1 binding in the brain, taking the allele information and quantile normalized DNase-seq profiles in 13 REMC fetal brain samples as input. Our model predicted differential binding of SP1 at the two alleles of rs12225399, showing higher predicted probability of binding at the C allele than the G allele across all 13 REMC samples (Figure 4.4B).

4.4 Discussion

We have proposed *PHLDB1* to be a candidate target gene repressed in the risk genotype of rs12803321. Our identified causal SNP rs12225399 also appears as one of top candidate causal SNPs in a previous

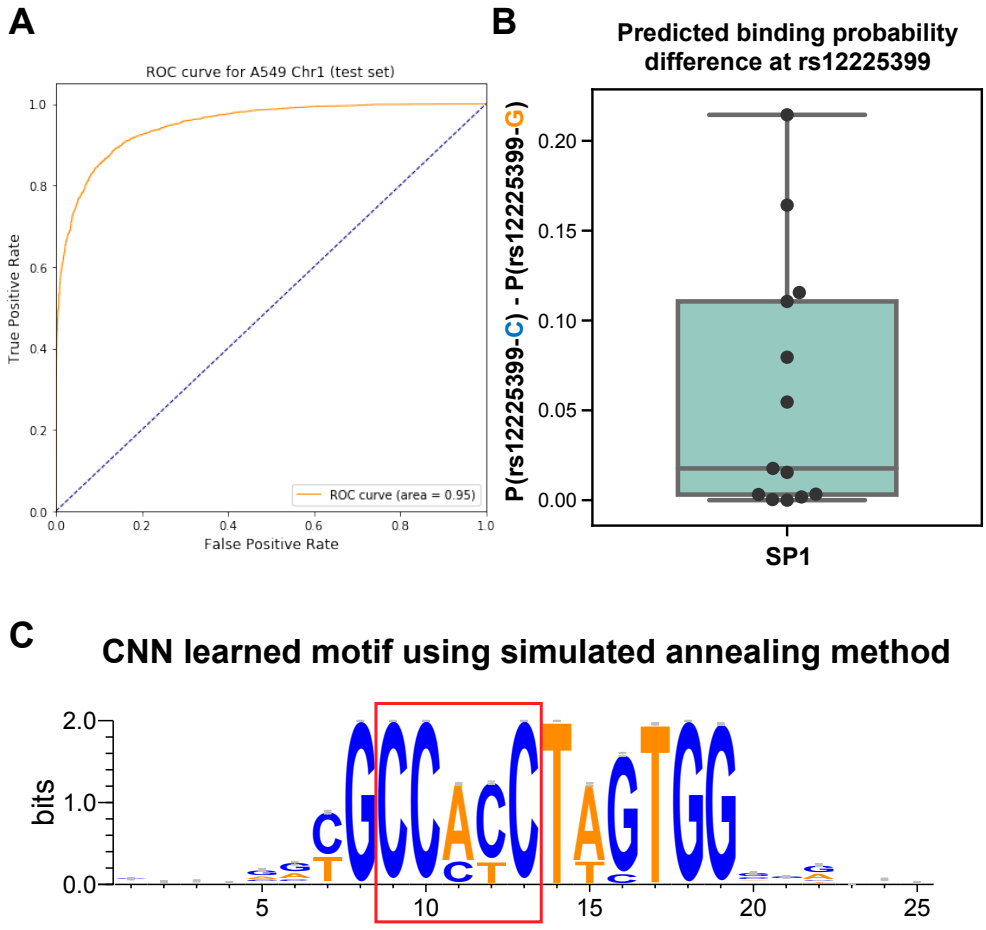


Figure 4.4: The CNN model predicted higher binding probability of SP1 at the rs12225399-C allele than the rs12225399-G allele. (A) Receiver operating characteristic (ROC) curve assessing the performance of the CNN model trained on six ENCODE SP1 ChIP-seq datasets (H1-hESC, HEK293T, HepG2, Liver, K562, MCF-7). Test set was chr1 data in the A549 cell line. Area under the curve (AUC) = 0.95. (B) The difference of SP1 binding probability between the two alleles of rs12225399, predicted by the CNN model based on 13 REMC fetal brain DNase-seq datasets from 10 donors. (C) CNN-learned motif visualized through a motif logo obtained from WebLogo [79] 3. The core motif inside the red box resembles the core motif of SP1 MA0079.3 (Figure 4.3A).

study implicating *PHLDB1* for a different GWAS SNP [27]. Knockdown of *PHLDB1* has been shown to increase cell death and reduce neurosphere formation in the U87MG glioma cell line [27], but its molecular function remains poorly understood. We have developed a deep learning approach for predicting the binding pattern of TFs when their ChIP-seq data are not available in the human brain. Most previous machine learning approaches have been using only sequence information for predicting protein binding patterns [32], and some recent studies have begun to utilize other genomic and epigenomic information [108]. Our deep learning model integrates DNase-seq signal with sequence information into one convolutional filter. Using

the convolutional neural network trained on non-brain cell data to evaluate sequence and open chromatin information in brain tissues has allowed us to predict allelic preference of SP1 binding. A similar approach may benefit future functional genomics studies in the brain, where TF ChIP-seq data are not readily available.

Chapter 5

The functional role of 3q14.1 variant rs11706832 - *LRIG1* locus

In this chapter, we present the functional role of 3q14.1 GWAS variant rs11706832. The work in this chapter has been published in Manjunath†, Yan† [1] *et al.*, Neuro-oncology, 2020 [51]. The analysis in this chapter was mainly done by the author.

5.1 eQTL and phased ASE analyses implicate *SLC25A26* as a candidate target gene

The LGG GWAS SNP rs11706832 (reference allele: A, alternative allele: C (risk)), located in an intron of Leucine rich repeats and immunoglobulin like domains 1 (*LRIG1*) (Figure 5.1), was reported to be associated with "IDH^{mut} only" and triple-positive glioma subgroups [104]. Although highly expressed in the brain, *LRIG1* did not show a significant eQTL association with rs11706832 in TCGA LGG data ($P = 0.52$ and 0.34 for "IDH^{mut}only" and triple-positive, respectively), in agreement with a previous report [12]. By contrast, we found that Solute carrier family 25 member 26 (*SLC25A26*), a gene 432kb away from *LRIG1*, was significantly associated with rs11706832 in eQTL and phased ASE analyses: the number of rs11706832 risk allele C was positively correlated with the expression level of *SLC25A26* (Figure 5.2; genotype $P = 2.9\times10^{-3}$, "IDH^{mut} only"; 4.11×10^{-2}, triple-positive; 2.11×10^{-4}, "IDH^{mut} only" and triple-positive combined; FDR $= 1.48 \times 10^{-3}$, "IDH^{mut} only" and triple-positive combined). Phased ASE analysis identified seven exonic SNPs with a Wilcoxon signed-rank sum test $P < 0.05$ ("IDH^{mut} only" and triple-positive combined group, case number > 5). Five of these 7 exonic SNPs showed a significant transcriptional skew toward the rs11706832-C allele (Figure 5.3), in agreement with the eQTL result, while the other two showed an opposite trend. These results suggested that a functional consequence of the GWAS risk allele rs11706832-C was to increase the expression of *SLC25A26*.

[1] †co-first authors, contributed equally.

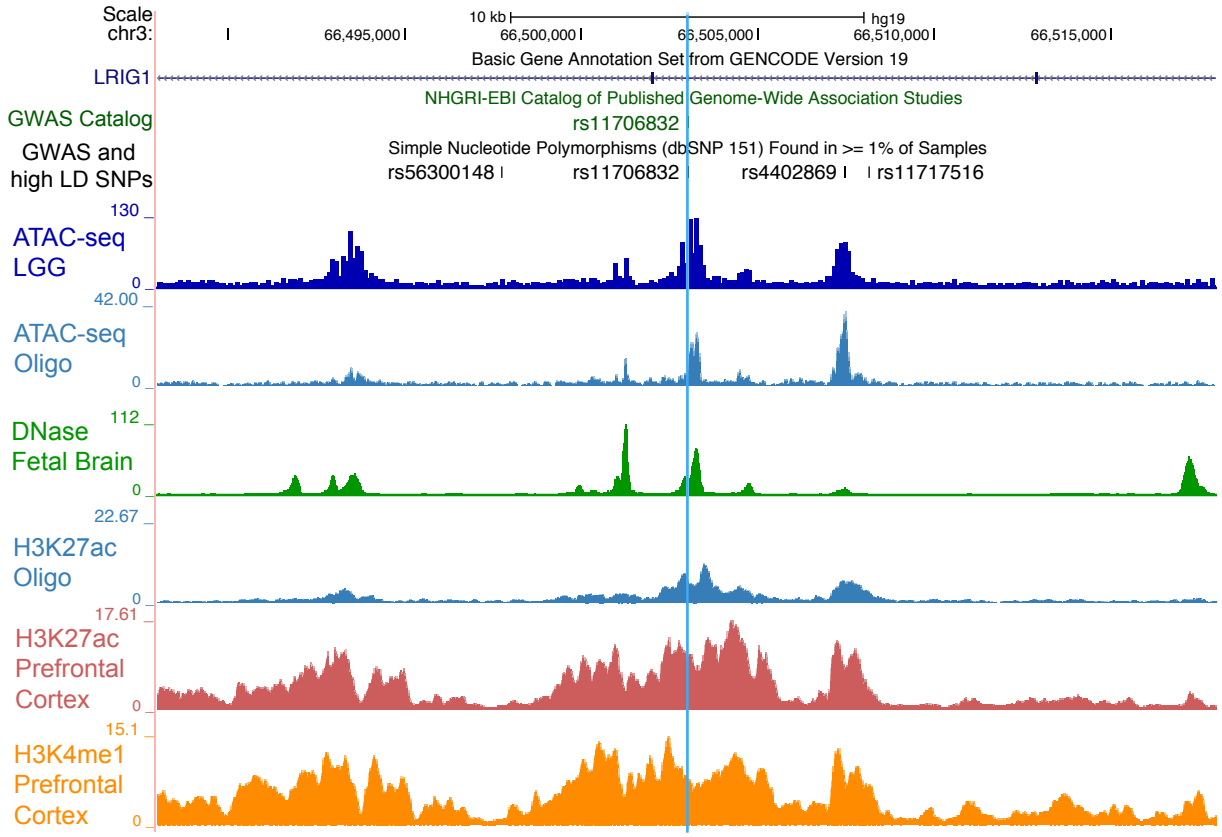


Figure 5.1: The GWAS SNP rs11706832 resides in a regulatory element within an *LRIG1* intron. A snapshot of the *LRIG1* locus where the GWAS SNP rs11706832 is denoted by a blue vertical line. The top three tracks are: basic gene annotation set from GENCODE version 19, LGG GWAS SNPs from GWAS catalog [106] and SNPs in high LD with rs11706832 (dbSNP [107] 151). The lower epigenomic tracks are: TCGA-LGG ATAC-seq [54], oligodendrocytes ATAC-seq [55], oligodendrocytes H3K27ac [55] and REMC data in fetal brain and prefrontal cortex.

5.2 Functional analysis of rs11706832 locus identifies the (rs11706832, *SLC25A26*, LEF1) triplet

Of all three SNPs in high LD with rs11706832 (Table D.1), rs4402869 ($r^2 = 0.87$) and the GWAS SNP rs11706832 resided in open chromatin and active enhancer regions (Figure 5.1). Motif analysis and gene-TF expression correlation analysis for rs11706832 and rs4402869 identified rs11706832-LEF1 to be the best candidate SNP-TF pair (Appendix D.3, D.4), with the rs11706832-A allele potentially creating a LEF1 binding motif (FIMO $P = 9.4 \times 10^{-4}$, Figure D.1A) and the A-to-C conversion significantly perturbating the binding motif (permutation test $P = 0.012$). The correlation structure between *LEF1* and *SLC25A26* expression was inconclusive in the combined "IDH^{mut} only" and triple-positive group, but the anticorrelation

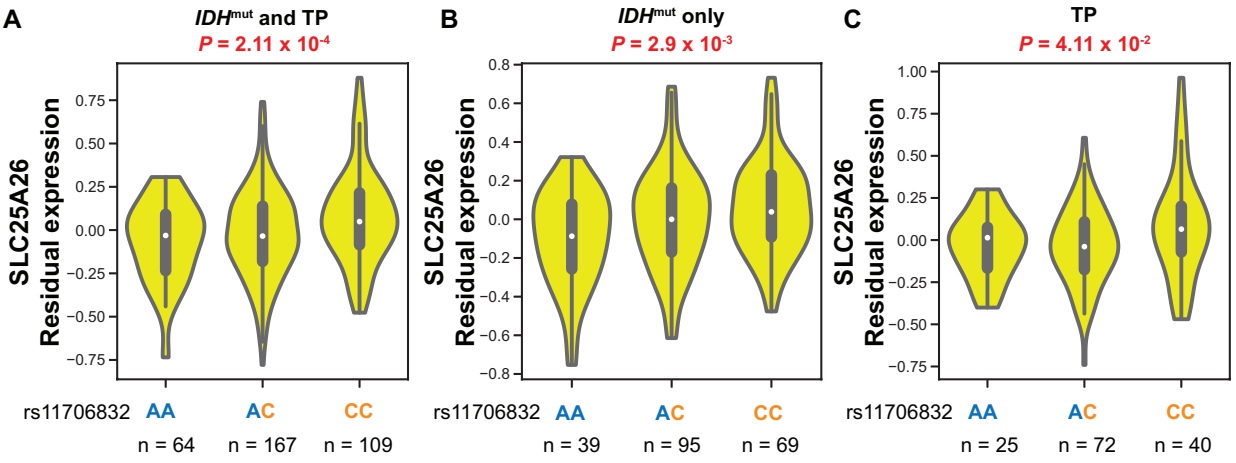


Figure 5.2: The GWAS SNP rs11706832 is associated with an increased expression of *SLC25A26*. (A) eQTL result for rs11706832 and *SLC25A26* in the combined TCGA-LGG "IDH^{mut} only" and triple positive (TP) group. (B) Similar to (A), but for the "IDH^{mut} only" subgroup. (C) Similar to (A), but for the triple positive (TP) subgroup.

was clearly strongest in the AA genotype when all LGG samples were used (Figure D.1B,C). These results together suggested that the rs11706832-A allele might create a binding site of LEF1, a known transcriptional repressor [109], thereby suppressing the expression of *SLC25A26*.

5.3 Discussion

From Section 5.1 and 5.2, we have shown *SLC25A26* expression to be elevated in the risk group. It is worth noting that this gene belongs to the mitochondrial carrier family and encodes a protein involved in transporting S-adenosylmethionine into the mitochondria [110]. It has been shown that overexpression of *SLC25A26* in CaSki cells contributes to mitochondrial DNA (mtDNA) hypermethylation [111] and that mtDNA methylation level tends to decrease during glioblastoma progression [112]. Future studies may reveal how potential mtDNA methylation changes attributable to *SLC25A26* modulation by rs11706832 contribute to LGG tumorigenesis.

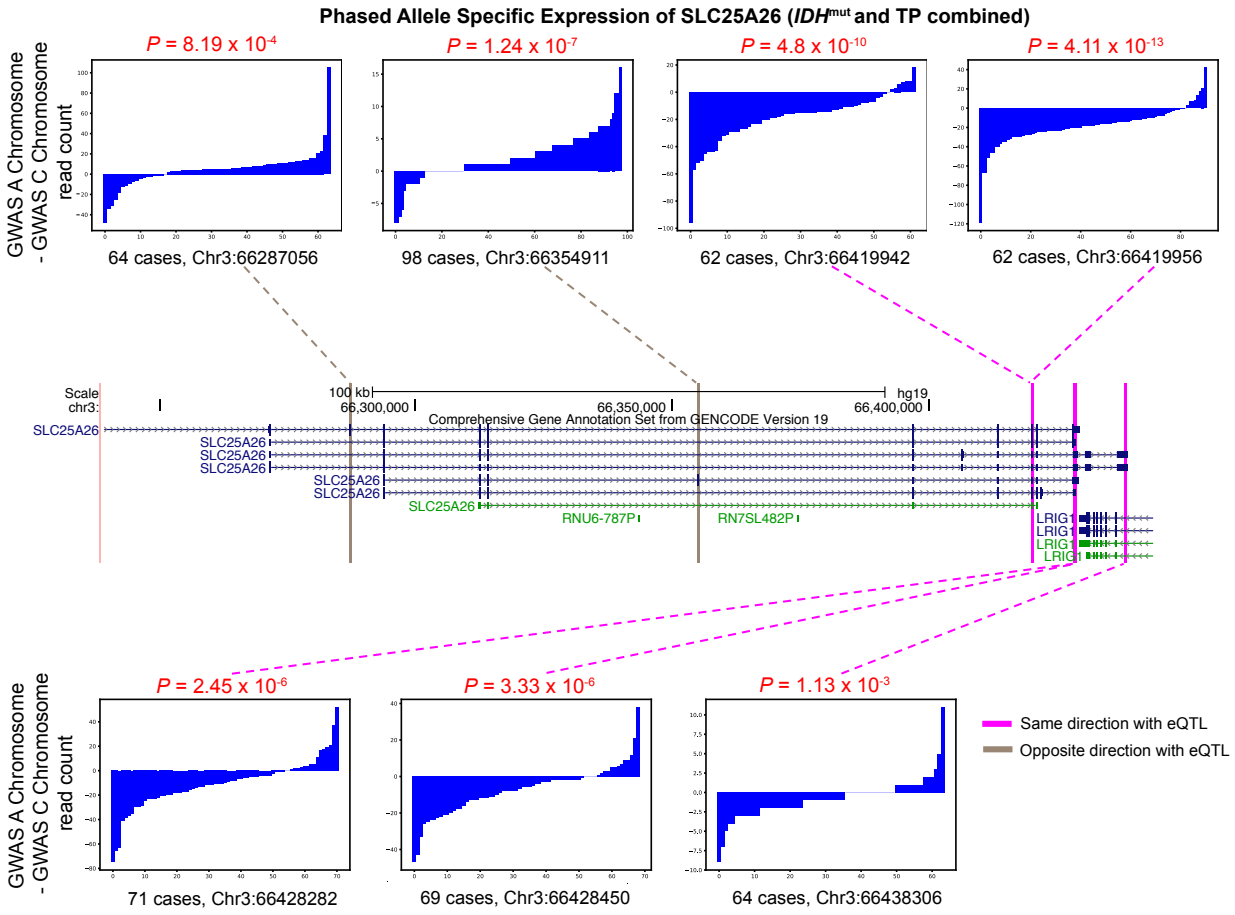


Figure 5.3: Phased allele-specific transcription pattern of *SLC25A26* at 7 exonic SNPs in the combined TCGA-LGG "IDH^{mut} only" and triple-positive group. 5 of these 7 exonic SNPs show a significant transcriptional skew toward the rs11706832-C allele (marked by magenta lines), in agreement with the eQTL result, while the other two show an opposite trend (marked by brown lines). For each exonic SNP, RNA-Seq read count differences between the two chromosomes harboring the rs11706832-A allele and the rs11706832-C allele are sorted across patients and shown as bar plots. The P-value from the Wilcoxon signed-rank sum test is shown at the top of each bar plot. The genomic locations of these 7 exonic SNPs are shown in the GENCODE Version 19 track.

Chapter 6

Neural network parameter reduction using tensor train decomposition

In this chapter, we first review the related concepts of TT-decomposition and the idea of applying TT-decomposition to neural network parameter reduction. We then demonstrate that the parameter-reduced CNN of SP1 performs well. The analysis in this chapter was performed by the author. The results in this chapter have not been published.

6.1 Tensor train decomposition (TT-decomposition)

Tensors can be viewed as multidimensional generalizations of vectors and matrices. Tensor is defined as a d-dimensional array: $\mathbf{A} = A(i_1, \ldots, i_d)$, where $\mathbf{A}$ denotes a tensor and i_k ($k = 1, \ldots, d$; $1 \leqslant i_k \leqslant n_k$) denotes the dimension index. In Chapter 1, we presented two ways for performing tensor decomposition: canonical polyadic decomposition (CP decomposition) and higher-order singular value decomposition (HOSVD). Both of them have certain disadvantages, which motivates us to study a new format of tensor decomposition, called tensor train decomposition (TT-decomposition). We will explain the TT-format of a given tensor in detail in this section.

6.1.1 TT-format

Suppose we approximate a given tensor $\mathbf{B}$ by a tensor $\mathbf{A}$ ($\mathbf{A} \approx \mathbf{B}$), whose elements are expressed as [34]:

$$A(i_1, \ldots, i_d) = G_1(i_1)G_2(i_2)\ldots G_d(i_d), \tag{6.1}$$

where $G_k(i_k)$ is an $r_{k-1} \times r_k$ matrix [34]. We call the matrix product in equation 6.1 the tensor train format (TT-format) of tensor $\mathbf{A}$. Equation 6.1 can be interpreted as: one tensor element with index $(i_1, \ldots, i_d)$ is expressed by the product of d matrices, where the k^{th} ($1 \leqslant k \leqslant d$) matrix has dimension $r_{k-1} \times r_k$, with $r_0 = 1$ and $r_d = 1$. Denoting n_k as the number of all available values of i_k, the above mentioned $G_k(i_k)$ is actually a $r_{k-1} \times n_k \times r_k$ three-dimensional array, with elements $G_k(\beta_{k-1}, i_k, \beta_k) = G_k(i_k)_{\beta_{k-1}\beta_k}$. Now we

can write equation 6.1 in the index form [34]:

$$A(i_1, \ldots, i_d) = \sum_{\beta_0 \ldots \beta_d} G_1(\beta_0, i_1, \beta_1) G_2(\beta_1, i_2, \beta_2) \ldots G_d(\beta_{d-1}, i_d, \beta_d). \tag{6.2}$$

To better illustrate the tensor train format, we give an example in Figure 6.1. The tensor element A_{2132} is equal to the product of the 4 matrices marked by blue. Each core G_k ($k \in [1, 2, 3, 4]$) represents an $r_{k-1} \times n_k \times r_k$ three-dimensional array, where r_k are the compression ranks or TT-ranks of the tensor train decomposition, and n_k is the dimension of i_k. The TT-format uses $O(ndr^2)$ memory to store n^d elements in the tensor, which is efficient when the TT-ranks are small.

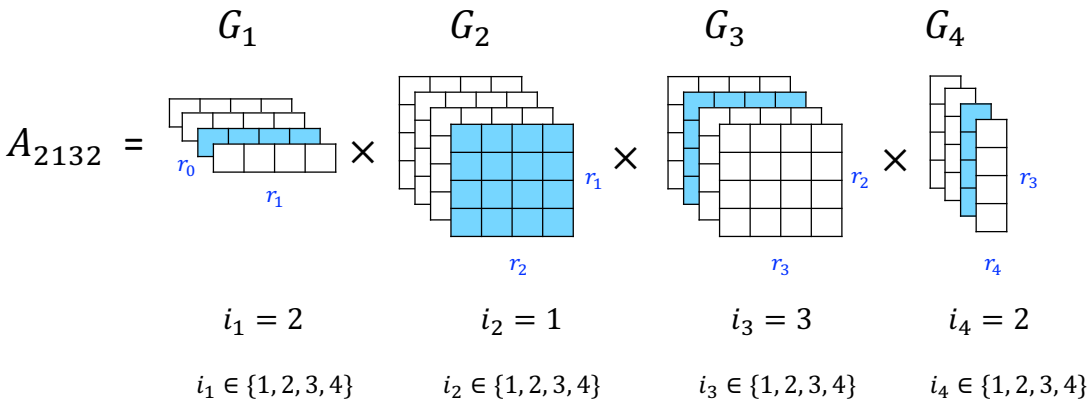


Figure 6.1: An example of tensor train decomposition. The figure is modified from https://t3f.readthedocs.io/en/latest/faq.html [113]. A_{2132} denotes one element in tensor $\mathbf{A}$ whose indexes are: $i_1 = 2$, $i_2 = 1$, $i_3 = 3$ and $i_4 = 2$; G_k ($k \in [1, 2, 3, 4]$) denote the cores of TT-decomposition, and the matrices whose product is equal to A_{2132} are marked by blue; r_k denote the ranks of the TT-decomposition.

We could also express the tensor train format in the following graphical way [114, 115]. For example, when $d = 4$, the graphical representation of TT-format is:

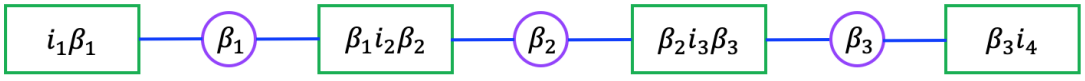


Figure 6.2: Graphical representation of the tensor train format, figure modified from I. V. Oseledets [34].

There are two types of nodes in Figure 6.2: the rectangular node and the circle node [34]. Rectangular nodes contain spatial indices i_k and auxiliary indices β_k [34]. Circles only have the auxiliary indices β_k and represent a link: we connect the two cores if an auxiliary index is present in these two cores [34]. Here, we also assume the summation over the auxiliary indices, which corresponds to the summation in equation 6.2 [34]. The graphical representation looks like a train, which corresponds to the name "tensor train decomposition". It is worth to mention that this name was first proposed by I. V. Oseledets.

6.1.2 TT-SVD algorithm

We have presented the TT-format in Section 6.1.1. The question arises next is how to get the TT-format either exactly or approximately for a tensor. Mathematically, we could describe the approximation question as below: given a prescribed accuracy ε, how to approximate a given tensor $\mathbf{A}$ with $\hat{\mathbf{A}}$ in the TT-format such that $\|\mathbf{A} - \hat{\mathbf{A}}\|_F \leq \varepsilon \|\mathbf{A}\|_F$ [34]?

I. V. Oseledets first proposed the TT-SVD algorithm to obtain the TT-format of a given tensor [34]. As implicated by the name, TT-SVD algorithm computes the TT-decomposition through d sequential singular value decompositions (SVDs) [34]. In this section, we will first give the definition of the unfolding matrix, which is heavily used in the algorithm, then present the stepwise TT-SVD algorithm.

The unfolding matrix of tensor $\mathbf{A}$ is defined as:

$$A_k = A_k(i_1, \ldots, i_k; i_{k+1}, \ldots, i_d) = A(i_1, \ldots, i_d), \tag{6.3}$$

where the first k indices enumerate the rows of A_k, and the last $d-k$ indices enumerate the columns of A_k [34]. We can think this unfolding as: for a element of $\mathbf{A}$ indexed by $(i_1, \ldots, i_d)$, its position in the unfolded matrix is row $(i_1, \ldots, i_k)$ and column $(i_{k+1}, \ldots, i_d)$. From the definition, we could see that the size of the matrix is $(\prod_{s=1}^{k} n_s) \times (\prod_{s=k+1}^{d} n_s)$ [34].

Suppose the unfolding matrix of a given tensor is of low rank only approximately, i.e., the unfoldings A_k of the tensor $\mathbf{A}$ satisfy: $A_k = R_k + Q_k$, where $\operatorname{rank}(R_k) = r_k$, $\|Q_k\|_F = \varepsilon_k$ $(k = 1, \ldots, d-1)$ [34]. It can be shown [34] that the TT-SVD algorithm computes a tensor $\hat{\mathbf{A}}$ in the TT-format with TT-ranks r_k and

$$\|\mathbf{A} - \hat{\mathbf{A}}\|_F \leq \sqrt{\sum_{k=1}^{d-1} \varepsilon_k^2}. \tag{6.4}$$

With the unfolding matrix defined and the above theorem stated, we then present the TT-SVD algorithm below. The TT-SVD algorithm starts from the truncation parameter δ, which is equal to $\frac{\varepsilon}{\sqrt{d-1}}\|\mathbf{A}\|_F$, where

ε is the given accuracy. The algorithm also involves reshaping the matrices and computing the δ-truncated SVD iteratively.

Algorithm 2 TT-SVD algorithm, modified slightly from I. V. Oseledets [34].

Given: A d dimensional tensor $\mathbf{A}$, accuracy ε.
Goal: Get the cores $G_1, \ldots, G_d$ of the TT-approximation $\hat{\mathbf{A}}$ to $\mathbf{A}$ with TT-ranks $\hat{r_k}$ of $\hat{\mathbf{A}}$ equal to the δ-ranks of the unfoldings A_k of $\mathbf{A}$, where $\delta = \frac{\varepsilon}{\sqrt{d-1}}\|\mathbf{A}\|_F$. $\hat{\mathbf{A}}$ satisfies $\|\mathbf{A} - \hat{\mathbf{A}}\|_F \leq \varepsilon\|\mathbf{A}\|_F$.
Begin: Compute the truncation parameter $\delta = \frac{\varepsilon}{\sqrt{d-1}}\|\mathbf{A}\|_F$. Set the temporary tensor as $\mathbf{D}$, $\mathbf{D} = \mathbf{A}$. $r_0 = 1$.
for k **in** $(1, 2, \ldots, d-1)$ **do**
    $D := \text{reshape}(D, [r_{k-1}n_k, \frac{The\ number\ of\ elements\ in\ D}{r_{k-1}n_k}])$
    Compute δ-truncated SVD: $D = USV^T + E, \|E\|_F \leq \delta, r_k = rank_\delta(D)$
    Get the k^{th} core: $G_k = \text{reshape}(U, [r_{k-1}, n_k, r_k])$
    Set $D = SV^T$
end for
$G_d = D$
Return: Tensor $\mathbf{A}$'s approximation $\hat{\mathbf{A}}$ in the TT-format: $G_1 G_2 \ldots G_d$.

6.2 TT-decomposition applied to neural network compression - tensor net

In a convolutional neural network, the fully connected layer transforms a high-dimensional input signal (denoted by $\boldsymbol{x}$) to a high-dimensional output signal (denoted by $\boldsymbol{y}$) with a large dense matrix $\boldsymbol{W}$ and bias vector $\boldsymbol{b}$:

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}. \tag{6.5}$$

The parameter number in the fully connected layer often consists of the largest portion of the total parameter number in a convolutional neural network, as a result of which large training datasets are required and overfitting could potentially occur. Therefore, numerous attempts have been made to reduce the parameter number of the fully connected layer. For example, studies utilizing the low-rank approximation of the weight matrix have been performed [116, 117, 40]. In line with the low-rank idea, A. Novikov *et al.* [40] proposed to treat the weight matrix as a multi-dimensional tensor and employed TT-format [34] to represent the tensor. One of the advantages of this method is that the back-propagation [118] derivatives can be computed using the properties of the TT-format [34, 40]. Following the naming convention proposed by A. Novikov *et al.* [40], we refer to the fully-connected layer in the TT-format as "TT-layer" and the neural network with one or more TT-layers as "tensor net". In the remaining part of this section, we first introduce the TT-

representation of vectors and matrices [34, 40], then introduce the TT-format of the linear transformation in a fully-connected layer [40], and at last present the learning process of the TT-layer embedded neural network [40].

Consider a vector $\boldsymbol{b}$ with length $N = \prod_{k=1}^{d} n_k$. We can always construct a bijection from $l \in \{1, 2, \ldots, N\}$ to the d-dimensional index vector $\boldsymbol{\mu}(l) = (\mu_1(l), \mu_2(l), \ldots, \mu_d(l))$, where $\mu_k(l) \in \{1, 2, \ldots n_k\}$. With the bijection defined, we can write the vector in the form of a d-dimensional tensor $\mathbf{B}$:

$$b(l) = B(\mu_1(l), \mu_2(l), \ldots, \mu_d(l)). \tag{6.6}$$

Thus, the elements in vector $\boldsymbol{b}$ is stored in a tensor $\mathbf{B}$. A vector in the TT-format is called a TT-vector [40].

Similarly, for a matrix $\boldsymbol{W}$ with dimension $M \times N$, where $M = \prod_{k=1}^{d} m_k$, $N = \prod_{k=1}^{d} n_k$, we can construct the bijection from both row index t and column index l to vectors: $\boldsymbol{\tau}(t) = (\tau_1(t), \tau_2(t), \ldots, \tau_d(t))$, $\boldsymbol{\mu}(l) = (\mu_1(l), \mu_2(l), \ldots, \mu_d(l))$, where $\tau_k(t) \in \{1, \ldots, m_k\}$ and $\mu_k(l) \in \{1, \ldots, n_k\}$. Thus, we store the matrix as a d-dimensional tensor $\mathbf{W}$, whose k^{th} dimension is of length $m_k n_k$, and is indexed by a long index $(\tau_k(t), \mu_k(l))$ [34, 40]:

$$W(t, l) = W(((\tau_1(t), \mu_1(l)), \ldots, (\tau_d(t), \mu_d(l))). \tag{6.7}$$

The TT-format of $\boldsymbol{W}$ with dimension $M \times N$ is thus [40]:

$$\begin{aligned} W(t, l) &= W(((\tau_1(t), \mu_1(l)), \ldots, (\tau_d(t), \mu_d(l))) \\ &= G_1[\tau_1(t), \mu_1(l)] \ldots G_d[\tau_d(t), \mu_d(l)]. \end{aligned} \tag{6.8}$$

The matrices $G_k[\tau_k(t), \mu_k(l)]$ $(k = 1, \ldots, d)$ are the TT-cores with tuple $(\tau_k(t), \mu_k(l))$ being an index. A matrix in the TT-format is called a TT-matrix [40].

Now we can write down the TT-format of the linear transformation $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}$ ($\boldsymbol{W} \in \mathbb{R}^{M \times N}$, $\boldsymbol{b} \in \mathbb{R}^{M}$) in a fully connected layer as [40]:

$$Y(i_1, \ldots, i_d) = \sum_{j_1, \ldots, j_d} G_1[i_1, j_1] \ldots G_d[i_d, j_d] X(j_1, \ldots, j_d) + B(i_1, \ldots, i_d). \tag{6.9}$$

In equation 6.9, the matrix is written in the TT-format, while the vector $\boldsymbol{x}$ is stored explicitly as a d-dimentional tensor $\mathbf{X}$, i.e., we do not decompose $\mathbf{X}$ into the product of TT-cores $X_1(j_1) \ldots X_d(j_d)$. The computational complexity for the TT-matrix-by-explicit-vector product $\boldsymbol{z} = \boldsymbol{W}\boldsymbol{x} = \sum_{j_1, \ldots, j_d} G_1[i_1, j_1] \ldots G_d[i_d, j_d]$

$X(j_1, \ldots, j_d)$ is $O(dr^2 m \max\{M, N\})$ [40], where d is the number of the cores of the TT-matrix; r is the maximum of ranks r_k $(k = 1, \ldots, (d-1))$; m is the maximum of m_k $(k = 1, \ldots, d)$; $M = \prod_{k=1}^{d} m_k$ is equal to the row number of the weight matrix $\boldsymbol{W}$; and $N = \prod_{k=1}^{d} n_k$ is equal to the column number of the weight matrix $\boldsymbol{W}$.

To update all the trainable parameters in the neural network learning process, we need to calculate the gradient of the loss function L with regards to the parameters through the back-propagation procedure [118]. Suppose $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{W}$ and $\boldsymbol{b}$ are the input vector, output vector, weight matrix and bias vector of the fully connected layer, the back-propagation procedure computes the following derivatives:

$$\frac{\partial L}{\partial \boldsymbol{W}} = \frac{\partial L}{\partial \boldsymbol{y}} \boldsymbol{x}^T, \qquad \frac{\partial L}{\partial \boldsymbol{b}} = \frac{\partial L}{\partial \boldsymbol{y}}, \qquad \frac{\partial L}{\partial \boldsymbol{x}} = \boldsymbol{W}^T \frac{\partial L}{\partial \boldsymbol{y}}. \tag{6.10}$$

We are able to update the the bias vector $\boldsymbol{b}$ in the fully connected layer and the parameters in previous layers using the derivatives given in equation 6.10, where $\frac{\partial L}{\partial \boldsymbol{x}} = \boldsymbol{W}^T \frac{\partial L}{\partial \boldsymbol{y}}$ is calculated by the TT-matrix-by-explicit-vector product [40]. In order to update the weight parameters, one way is to compute the gradient of the loss function with regards to the weight matrix, then convert the gradient matrix into the TT-format using the TT-SVD algorithm [34], and then add the gradient multiplied by a step size to the current weight matrix: $\boldsymbol{W}_{iter+1} = \boldsymbol{W}_{iter} + \gamma_{iter} \frac{\partial L}{\partial \boldsymbol{W}}$ [40]. However, this approach requires $O(MN)$ memory to compute $\frac{\partial L}{\partial \boldsymbol{W}}$ [40]. Thus, A. Novikov *et al.* proposed to compute the gradient of the loss function with regards to the cores of the TT-matrix directly [40]. In order to make the notation clear, A. Novikov *et al.* used the compressed notation for the indices before and after i_k: $\boldsymbol{i}_k^- := (i_1, \ldots, i_{k-1})$, $\boldsymbol{i}_k^+ := (i_{k+1}, \ldots, i_d)$, $\boldsymbol{i} = (\boldsymbol{i}_k^-, i_k, \boldsymbol{i}_k^+)$ [40], and wrote the partial core products as [40]:

$$\boldsymbol{P}_k^-[\boldsymbol{i}_k^-, \boldsymbol{j}_k^-] := G_1[i_1, j_1] \ldots G_{k-1}[i_{k-1}, j_{k-1}], \quad \boldsymbol{P}_k^+[\boldsymbol{i}_k^+, \boldsymbol{j}_k^+] := G_{k+1}[i_{k+1}, j_{k+1}] \ldots G_d[i_d, j_d]. \tag{6.11}$$

Equation 6.9 for k $(k = 2, \ldots, d-1)$ was thus written as [40]:

$$Y(\boldsymbol{i}) = Y(\boldsymbol{i}_k^-, i_k, \boldsymbol{i}_k^+) = \sum_{\boldsymbol{j}_k^-, j_k, \boldsymbol{j}_k^+} \boldsymbol{P}_k^-[\boldsymbol{i}_k^-, \boldsymbol{j}_k^-] G_k[i_k, j_k] \boldsymbol{P}_k^+[\boldsymbol{i}_k^+, \boldsymbol{j}_k^+] X(\boldsymbol{j}_k^-, j_k, \boldsymbol{j}_k^+) + B(\boldsymbol{i}). \tag{6.12}$$

The gradient of the loss function L with regards to the k^{th} core in the position $[\tilde{i}_k, \tilde{j}_k]$ can be written as [40]:

$$\begin{aligned}\frac{\partial L}{\partial G_k[\tilde{i}_k, \tilde{j}_k]} &= \sum_{\boldsymbol{i}} \frac{\partial L}{\partial Y(\boldsymbol{i})} \frac{\partial Y(\boldsymbol{i})}{\partial G_k[\tilde{i}_k, \tilde{j}_k]} \\ &= \sum_{\boldsymbol{i}} \frac{\partial L}{\partial Y(\boldsymbol{i})} \frac{\partial Y(\boldsymbol{i}_k^-, \tilde{i}_k, \boldsymbol{i}_k^+)}{\partial G_k[\tilde{i}_k, \tilde{j}_k]} \\ &= \sum_{\boldsymbol{i}} \frac{\partial L}{\partial Y(\boldsymbol{i})} \Big(\sum_{\boldsymbol{j}_k^-, \boldsymbol{j}_k^+} (\boldsymbol{P}_k^-[\boldsymbol{i}_k^-, \boldsymbol{j}_k^-])^T (\boldsymbol{P}_k^+[\boldsymbol{i}_k^+, \boldsymbol{j}_k^+])^T X(\boldsymbol{j}_k^-, \tilde{j}_k, \boldsymbol{j}_k^+) \Big). \end{aligned} \tag{6.13}$$

For the gradient matrice

$$\frac{\partial Y(\boldsymbol{i})}{\partial G_k[\tilde{i}_k, \tilde{j}_k]} = \sum_{\boldsymbol{j}_k^-, \boldsymbol{j}_k^+} (\boldsymbol{P}_k^-[\boldsymbol{i}_k^-, \boldsymbol{j}_k^-])^T (\boldsymbol{P}_k^+[\boldsymbol{i}_k^+, \boldsymbol{j}_k^+])^T X(\boldsymbol{j}_k^-, \tilde{j}_k, \boldsymbol{j}_k^+), \tag{6.14}$$

one can utilize dynamic programming twice to compute [40]. The overall time complexity of the backward pass is $O(d^2 r^4 m \max\{M, N\})$, and the memory usage of the backward pass is $O(r^3 \max\{M, N\})$ [40].

For our experiments of SP1 binding prediction to be presented in Section 6.3, we utilized the package T3F, developed by A. Novikov *et al.* [113]. Since T3F uses TensorFlow as backend, we could incorporate the TT-layer into our original CNN model (Section 4.3) easily.

6.3 Parameter reduction of SP1 binding predictive model using tensor net

6.3.1 Motivation and the basic structure of the CNN-TT model

In Section 4.3, we presented a convolutional neural network model for TF SP1 binding prediction. There are 80141 parameters in the original CNN model, where the first fully connected layer consists of the majority (95.9%) of the total parameter number (Table 6.1). Therefore, to avoid overfitting, we translated the positive and negative datasets by -20bp, -10bp, 0bp, 10bp, 20bp, yielding 1029302 samples as the translated dataset for training and validation. However, it is difficult to obtain such large ChIP-seq dataset (even after translation) for other transcription factors. Moreover, it is a general trend to reduce the running time and memory requirement in the deep learning field. Recent studies have shown that the fully connected layer could be compressed without significant drop of the predictive power [116, 117, 119, 40]. We thus applied TT-decomposition [34, 40] to the first fully connected layer of the original CNN model, and tested its performance. We call the CNN model with the TT-layer embedded as the CNN-TT model in the following

context, to differentiate from the original CNN model in Section 4.3.

Layer	Number of parameters
Convolutional layer	2440
First dense layer	76880
Second dense layer	810
Third dense layer	11

Table 6.1: Parameter number of different layers in the original CNN model. First dense layer contains 95.9% of all parameters.

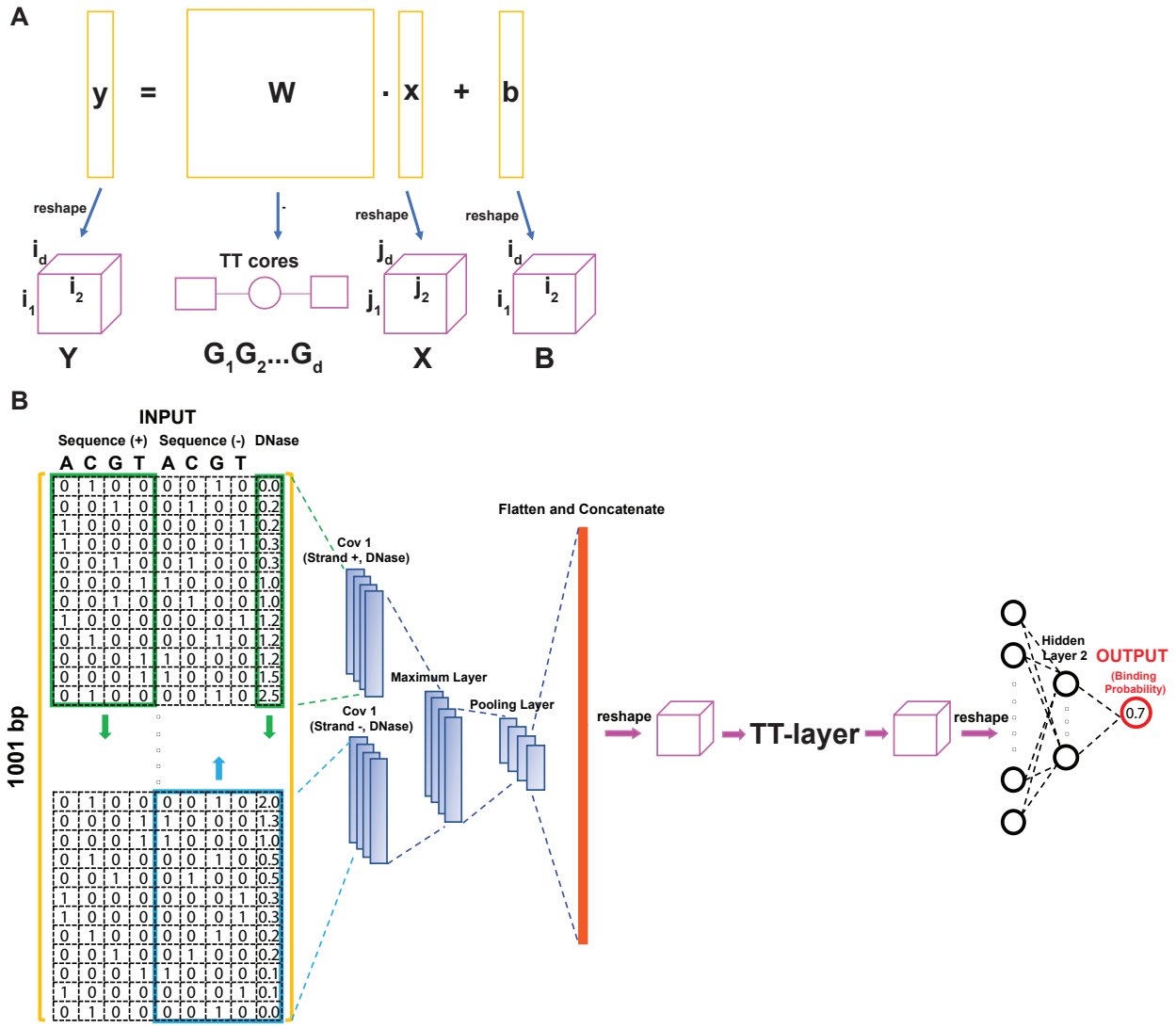


Figure 6.3: The basic structure of the TT-layer and the CNN-TT model. (A) The TT-format of the linear transformation $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}$, where the weight matrix was decomposed into the product of TT-cores $G_1 G_2 \dots G_d$, and the input $\boldsymbol{x}$, the output $\boldsymbol{y}$, the bias $\boldsymbol{b}$ were stored as tensor $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{B}$. The TT-format of the fully connected layer is thus: $Y(i_1, \dots, i_d) = \sum_{j_1,\dots,j_d} G_1[i_1, j_1] \dots G_d[i_d, j_d] X(j_1, \dots, j_d) + B(i_1, \dots, i_d)$. (B) The structure of the CNN-TT model for predicting the binding pattern of SP1 based on DNA sequence and open chromatin information. From left to right: 1001×9 input matrix incorporating sequence information and quantile-normalized DNase-seq signal at each base; convolutional layer using filters of length 12bp; maximum layer, extracting the maximum of the convolutional layer output from the positive and negative strands; maximum pooling layer; flatten and concatenate layer; TT-layer with input dimension 960 and output dimension 80; fully connected layer with 10 neurons; output.

The structure of the CNN-TT model is stated below. We substituted the first fully connected layer with the TT-layer, where the weight matrix was decomposed as the product of TT-cores $G_1 G_2 \dots G_d$, and the input vector $\boldsymbol{x}$, the output vector $\boldsymbol{y}$, the bias vector $\boldsymbol{b}$ were transformed to tensor $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{B}$ (Figure 6.3A). We kept other layers of the original CNN model unchanged, yielding the full structure of the CNN-TT model (Figure 6.3B). The training and validation dataset was the same with the dataset before translation for the original CNN model: there were 102934 samples in the positive set and 102934 samples in the negative set (from 6 cell types: H1-hESC, HEK293T, HepG2, Liver, K562, MCF-7; Section 4.3.1), yielding 205868 samples in total. We then split the 205868 samples into training and validation datasets with ratio 80% to 20% (training dataset: 164694 samples, validation dataset: 41174 samples). We trained the CNN-TT model using T3F library t3f.nn.KerasDense function [113], with different ranks of the TT-cores (Table 6.2). For each CNN-TT model configuration, the training was stopped after the validation loss reached plateau (Figure E.1), and the model was retrieved at the ending epoch. For each retrieved model, we tested its performance using SP1 A549 chromosome 1 positive and negative datasets (3785 samples for each dataset) and calculated the receiver operating characteristic (ROC) area under curve (AUC).

Rank distribution $(r_0, r_1, \dots, r_d)$	Weight matrix dimensions $(m_1, \dots, m_d) \times (n_1, \dots, n_d)$	Total parameter number	TT-layer parameter number	Percentage	ROC AUC on A549 chr1 test set
(1, 4, 4, 1)	(4, 4, 5) × (4, 4, 60)	4861	1600	2.1%	0.969
(1, 8, 8, 1)	(4, 4, 5) × (4, 4, 60)	6893	2688	3.5%	0.971

Table 6.2: The CNN-TT model predicted SP1 binding probability with high confidence, and the total parameter number was largely reduced from the original CNN model. Columns from left to right: the ranks $(r_0, r_1, \dots, r_d)$ of the TT-representation, where $r_0 = r_d = 1$; the weight matrix dimensions $(m_1, \dots, m_d)$ and $(n_1, \dots, n_d)$, where $M = \prod_{k=1}^{d} m_k$ and $N = \prod_{k=1}^{d} n_k$ are the total row and column numbers of the weight matrix, respectively; the total parameter number in each CNN-TT model; the parameter number in the TT-layer of each CNN-TT model; the percentage of the TT-layer parameter number to the original fully connected layer parameter number (76880); the receiver operating characteristic (ROC) area under curve (AUC) for each CNN-TT model.

6.3.2 The trained CNN-TT model predicts SP1 binding probability with high confidence

We tested the performance of the CNN-TT model using A549 chr1 positive and negative datasets (3785 samples for each). For the configurations listed in Table 6.2, the receiver operating characteristic (ROC) area under curve (AUC) are 0.969 and 0.971, respectively (Figure 6.4). We then applied the simulated-annealing method [69] to obtain the motifs learned by the CNN-TT models. Interestingly, we found that

apart from the SP1-like motif (Figure 6.5A, Figure E.2A, Figure 4.3A), the CNN-TT model also learned the motif of HNF4A (Figure 6.6, Figure E.3), a transcription co-factor of SP1 implicated in previous studies [120, 121]. These results together suggested that, although the total parameter number is largely reduced compared to the original CNN model, our CNN-TT model predicted SP1 binding probability with high confidence.

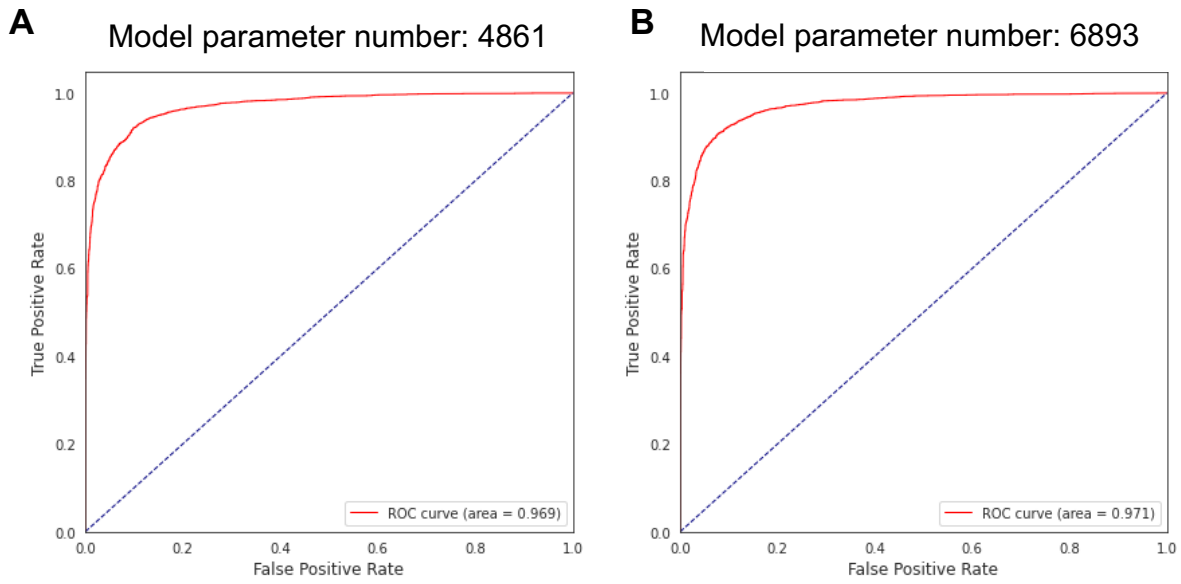


Figure 6.4: The trained CNN-TT models predict SP1 binding probability with high confidence in A549 chr1 test dataset. (A) The ROC curve of the CNN-TT model with configuration $(r_0, r_1, r_2, r_3) = (1, 4, 4, 1)$ and $(m_1, m_2, m_3) \times (n_1, n_2, n_3) = (4, 4, 5) \times (4, 4, 60)$. The total parameter number is 4861, and the AUC is 0.969. (B) Similar to (A), but for configuration $(r_0, r_1, r_2, r_3) = (1, 8, 8, 1)$ and $(m_1, m_2, m_3) \times (n_1, n_2, n_3) = (4, 4, 5) \times (4, 4, 60)$. The total parameter number is 6893, and the AUC is 0.971.

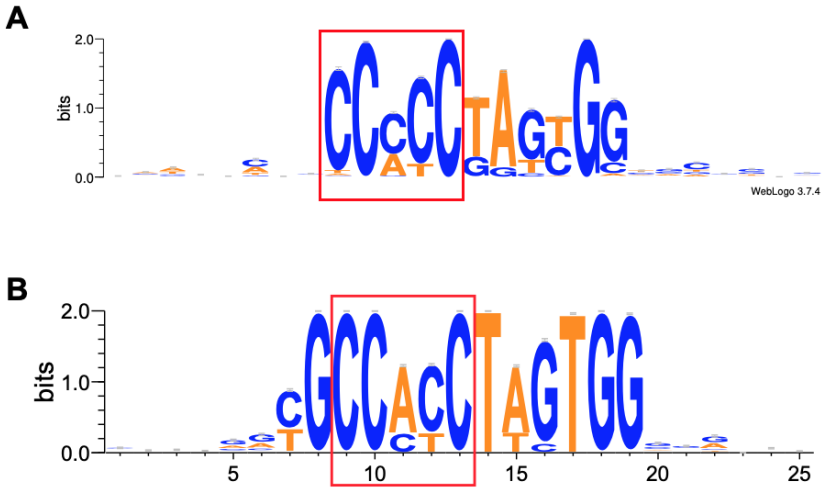


Figure 6.5: The core motif learned by the CNN-TT model resembled the core motif of SP1 MA0079.3. (A) One of the motifs learned by the CNN-TT model with configuration $(r_0, r_1, r_2, r_3) = (1, 4, 4, 1)$ and $(m_1, m_2, m_3) \times (n_1, n_2, n_3) = (4, 4, 5) \times (4, 4, 60)$ (total parameter number 4861), visualized through a motif logo obtained from WebLogo [79] 3. The core motif inside the red box resembles the core motif of SP1 MA0079.3 (Figure 4.3A). (B) The motif learned by the original CNN model in Section 4.3. The learned motif from the CNN-TT model resembles the learned motif from the original CNN model.

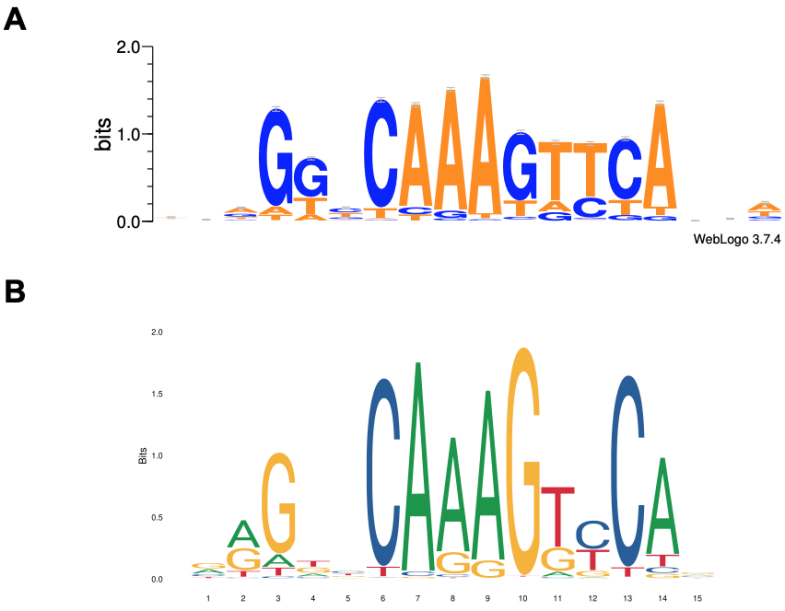


Figure 6.6: Another motif learned by the CNN-TT model resembles the HNF4A motif MA0114.2. (A) Another motif learned by the CNN-TT model with configuration $(r_0, r_1, r_2, r_3) = (1, 4, 4, 1)$ and $(m_1, m_2, m_3) \times (n_1, n_2, n_3) = (4, 4, 5) \times (4, 4, 60)$ (total parameter number 4861), visualized through a motif logo obtained from WebLogo [79] 3. (B) The motif logo of HNF4A MA0114.2 from the JASPAR [24] database.

6.4 Discussion

In this chapter, we reviewed the idea of using TT-decomposition to compress the fully connected layer in a neural network. The key for transforming the weight matrix $\boldsymbol{W}$ in a fully connected layer into the TT-format is to rearrange the weight matrix as a multi-dimensional tensor, and utilize TT-decomposition to approximate the tensor as a product of TT-cores. In the example of SP1 binding prediction, the ROC AUC on the test dataset is above 0.969 for the CNN-TT model configurations, where the parameter number of the TT-layer is as low as 2.1% of that in the original fully connected layer (Table 6.2). This suggested the parameters in the weight matrix of the fully connected layer are highly redundant, and it is possible for the neural network to capture the features of the input training data using a far more compact format. We applied simulated-annealing method [69] to extract the motifs learned, and visualized the extracted motifs using WebLogo [79] 3. Interestingly, we found that the CNN-TT model not only captured the motif which resembled the known SP1 motif, but also captured the motif of HNF4A (Figure 6.6, Figure E.3), a transcription co-factor of SP1 implicated in previous studies [120, 121]. These results together suggested that, the TT-layer embedded convolutional neural network could capture the data features in the training dataset with largely reduced parameter number; it relaxes the restriction of the large training dataset, and thus may have wider application in TF binding prediction.

Chapter 7

Conclusion

After entering the 21 century, the rapid development of the massive parallel sequencing technology has provided unprecedented opportunities for cancer research. It allowed us to unveil the function of the genome outside of the protein coding regions. With the help of various experimental methods such as ChIP-seq, we could obtain heterogeneous information ranging from transcription factor binding sites, histone modifications, to the 3D structure of the chromatin, all of which make the functional analysis of GWAS SNPs possible.

In this thesis, we first presented the framework for identifying the causative consequences of LGG GWAS SNPs. Based on the hypothesis that the GWAS loci contain causal SNPs that reside in functional regulatory regions of the human genome and modulate the expression of target genes by directly perturbing the binding affinity of TFs, we incorporated heterogeneous genomic, epigenomic and transcriptomic high-throughput sequencing data, and identified putative (causal SNP, target gene, TF) triplets. For 11q23.2 GWAS SNP rs648044, we have shown that rs648044 may modulate the expression of *ZBTB16* by perturbing the binding affinity of MAFF. We also proposed that *CIC*, one of the most commonly mutated genes in IDH^{mut} oligodendrogliomas and located on chromosome 19q, is likely a direct transcriptional target of ZBTB16. These observations suggested rs648044 might contribute to LGG pathogenesis through disrupting the potentially important regulation network involving ZBTB16 and *CIC*. For 11q23.3 GWAS SNP rs12803321, we identified *PHLDB1* as the putative target gene, and rs12225399, SP1/SP2 as the best candidate causal SNP, TF pair. For 3q14.1 GWAS SNP rs11706832, a similar approach revealed *SLC25A26*, a gene belongs to the mitochondrial carrier family and encodes a protein involved in transporting S-adenosylmethionine into the mitochondria, as the target gene, and (rs11706832, *SLC25A26*, LEF1) as the best candidate triplet.

Owing to the rapid development of the high performance computing cluster and the graphics processing unit (GPU), it has become approachable and convenient to apply deep learning method to prediction problems in the biological field. Deep learning approach like convolutional neural network permits the prediction models to learn the internal data structures and features. In our study, we developed a deep learning approach for predicting the binding pattern of TFs when their ChIP-seq data are not available in the human

brain. We integrated epigenomic information (DNase-seq signal) and genomic information (sequence) into one convolutional filter, and trained the CNN on non-brain cell data. Using the trained model to evaluate sequence and open chromatin information in brain tissues, we predicted the allele-specific binding preference for SP1 at rs12225399, consistent with our motif analysis. Furthermore, by utilizing the simulated-annealing-based optimization method, we confirmed the CNN-learned motif resembled the core binding motif of SP1. A similar approach may benefit future functional genomics studies in the brain, where TF ChIP-seq data are not readily available.

We at last presented the experiment of using TT-decomposition to compress the fully connected layer in a convolutional neural network. In the example of SP1 binding prediction, we demonstrated that high prediction accuracy was achieved using the TT-layer embedded CNN, where the parameter number of the TT-layer was as low as 2.1% of that in the original fully connected layer. This suggested that the TT-layer embedded convolutional neural network was able to capture the data features in the training dataset with largely reduced parameter number. It relaxes the restriction of the large training dataset required by the deep learning method in transcription factor binding prediction, and thus may have wider application in future functional genomics studies.

Although we identified candidate (causal SNP, target gene, TF) triplets that might contribute to LGG tumoregenesis in three case studies through the integrative computational framework, there were also certain limitations in our study. For example, even though our eQTL analysis took copy number alteration as a covariate in the linear model, it is possible that other uncharacterized somatic mutations that could alter transcription levels and mRNA stability might have strongly perturbed the mRNA abundance in tumor samples and complicated the target gene identification. Our study has focused on assessing the molecular function of genetic variants in altering the binding affinity of TFs, however, other molecular functions such as DNA methylation changes and protein modifications might also be important and need further investigation.

To facilitate the rapid identification of candidate (causal SNP, target gene, TF) triplets of the reported LGG GWAS SNPs, we have summarized our results into an interactive user-friendly web database, ALG3 (http://education.knoweng.org/alg3/), devoloped by Dr. Mohith Manjunath *et al.* [51]. As far as we are aware, there is no other work to date that presents the comprehensive characterization of the GWAS SNPs of LGG. We hope our proposed (causal SNP, target gene, transcription factor) triplets could facilitate additional analysis and expedite experimental validation. We also hope the physics-inspired deep learning methods provide insights to machine learning problems in the bioinformatics community.

Chapter 8

Reference

[1] R. D. Fields et al. "Glial biology in learning and cognition". In: *Neuroscientist* 20.5 (Oct. 2014), pp. 426–431.

[2] M. L. Goodenberger and R. B. Jenkins. "Genetics of adult glioma". In: *Cancer Genet* 205.12 (Dec. 2012), pp. 613–621.

[3] D. N. Louis et al. "The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary". In: *Acta Neuropathol* 131.6 (June 2016), pp. 803–820.

[4] J. E. Eckel-Passow et al. "Glioma Groups Based on 1p/19q, IDH, and TERT Promoter Mutations in Tumors". In: *N Engl J Med* 372.26 (June 2015), pp. 2499–2508.

[5] S. Tommasini-Ghelfi et al. "Cancer-associated mutation and beyond: The emerging biology of isocitrate dehydrogenases in human disease". In: *Sci Adv* 5.5 (May 2019), eaaw4543.

[6] T. Watanabe et al. "IDH1 mutations are early events in the development of astrocytomas and oligodendrogliomas". In: *Am J Pathol* 174.4 (Apr. 2009), pp. 1149–1153.

[7] Z. Turkalp, J. Karamchandani, and S. Das. "IDH mutation in glioma: new insights and promises for the future". In: *JAMA Neurol* 71.10 (Oct. 2014), pp. 1319–1325.

[8] D. Unruh et al. "Methylation and transcription patterns are distinct in IDH mutant gliomas compared to other IDH mutant cancers". In: *Sci Rep* 9.1 (June 2019), p. 8946.

[9] W. A. Flavahan et al. "Insulator dysfunction and oncogene activation in IDH mutant gliomas". In: *Nature* 529.7584 (Jan. 2016), pp. 110–114.

[10] G. M. Clarke et al. "Basic statistical analysis in genetic case-control studies". In: *Nat Protoc* 6.2 (Feb. 2011), pp. 121–133.

[11] G. S. Barsh et al. "Guidelines for genome-wide association studies". In: *PLoS Genet* 8.7 (July 2012), e1002812.

[12] B. S. Melin et al. “Genome-wide association study of glioma subtypes identifies specific differences in genetic susceptibility to glioblastoma and non-glioblastoma tumors”. In: *Nat Genet* 49.5 (May 2017), pp. 789–794.

[13] M. Levine. “Transcriptional enhancers in animal development and evolution”. In: *Curr Biol* 20.17 (Sept. 2010), R754–763.

[14] M. Bulger and M. Groudine. “Functional and mechanistic diversity of distal transcription enhancers”. In: *Cell* 144.3 (Feb. 2011), pp. 327–339.

[15] E. Calo and J. Wysocka. “Modification of enhancer chromatin: what, how, and why?” In: *Mol Cell* 49.5 (Mar. 2013), pp. 825–837.

[16] D. S. Johnson et al. “Genome-wide mapping of in vivo protein-DNA interactions”. In: *Science* 316.5830 (June 2007), pp. 1497–1502.

[17] I. Dunham et al. “An integrated encyclopedia of DNA elements in the human genome”. In: *Nature* 489.7414 (Sept. 2012), pp. 57–74.

[18] Y. Zhang et al. “SequencEnG: an interactive knowledge base of sequencing techniques”. In: *Bioinformatics* 35.8 (Apr. 2019), pp. 1438–1440.

[19] A. P. Boyle et al. “High-resolution mapping and characterization of open chromatin across the genome”. In: *Cell* 132.2 (Jan. 2008), pp. 311–322.

[20] J. D. Buenrostro et al. “Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position”. In: *Nat Methods* 10.12 (Dec. 2013), pp. 1213–1218.

[21] E. Lieberman-Aiden et al. “Comprehensive mapping of long-range interactions reveals folding principles of the human genome”. In: *Science* 326.5950 (Oct. 2009), pp. 289–293.

[22] R. Fang et al. “Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq”. In: *Cell Res* 26.12 (Dec. 2016), pp. 1345–1348.

[23] M. J. Fullwood et al. “An oestrogen-receptor-alpha-bound human chromatin interactome”. In: *Nature* 462.7269 (Nov. 2009), pp. 58–64.

[24] O. Fornes et al. “JASPAR 2020: update of the open-access database of transcription factor binding profiles”. In: *Nucleic Acids Res* 48.D1 (Jan. 2020), pp. D87–D92.

[25] Z. Tang et al. “CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription”. In: *Cell* 163.7 (Dec. 2015), pp. 1611–1627.

[26] I. Atkins et al. "Transcriptome-Wide Association Study Identifies New Candidate Susceptibility Genes for Glioma". In: *Cancer Res* 79.8 (Apr. 2019), pp. 2065–2071.

[27] R. Baskin et al. "Functional analysis of the 11q23.3 glioma susceptibility locus implicates PHLDB1 and DDX6 in glioma susceptibility". In: *Sci Rep* 5 (Nov. 2015), p. 17367.

[28] B. E. Bernstein et al. "The NIH Roadmap Epigenomics Mapping Consortium". In: *Nat Biotechnol* 28.10 (Oct. 2010), pp. 1045–1048.

[29] A. Kundaje et al. "Integrative analysis of 111 reference human epigenomes". In: *Nature* 518.7539 (Feb. 2015), pp. 317–330.

[30] R. L. Grossman et al. "Toward a Shared Vision for Cancer Genomic Data". In: *N Engl J Med* 375.12 (Sept. 2016), pp. 1109–1112.

[31] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[32] B. Alipanahi et al. "Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning". In: *Nat Biotechnol* 33.8 (Aug. 2015), pp. 831–838.

[33] D. Quang and X. Xie. "FactorNet: A deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data". In: *Methods* 166 (Aug. 2019), pp. 40–47.

[34] Ivan Oseledets. "Tensor-Train Decomposition". In: *SIAM J. Scientific Computing* 33 (Jan. 2011), pp. 2295–2317. DOI: 10.1137/090752286.

[35] Johan Håstad. "Tensor rank is NP-complete". In: *Journal of Algorithms* 11.4 (1990), pp. 644–654. ISSN: 0196-6774.

[36] Vin de Silva and Lek-Heng Lim. "Tensor Rank and the Ill-Posedness of the Best Low-Rank Approximation Problem". In: *SIAM Journal on Matrix Analysis and Applications* 30.3 (2008), pp. 1084–1127.

[37] L. R. Tucker. "Some mathematical notes on three-mode factor analysis". In: *Psychometrika* 31.3 (Sept. 1966), pp. 279–311.

[38] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. "A Multilinear Singular Value Decomposition". In: *SIAM Journal on Matrix Analysis and Applications* 21.4 (2000), pp. 1253–1278.

[39] Jun S. Song. *Physics 598 "Statistical Data Analysis and Stochastic Processes in Physics" lecture notes, Fall 2019*. Chap. 5.

[40] Alexander Novikov et al. “Tensorizing Neural Networks”. In: *Advances in Neural Information Processing Systems*. Vol. 28. Curran Associates, Inc., 2015, pp. 442–450.

[41] Yinchong Yang, Denis Krompass, and Volker Tresp. “Tensor-train recurrent neural networks for video classification”. In: *International Conference on Machine Learning*. PMLR. 2017, pp. 3891–3900.

[42] Jiahao Su et al. “Convolutional tensor-train lstm for spatio-temporal learning”. In: *arXiv preprint arXiv:2002.09131* (2020).

[43] J Ignacio Cirac and Frank Verstraete. “Renormalization and tensor product states in spin chains and lattices”. In: *Journal of Physics A: Mathematical and Theoretical* 42.50 (2009), p. 504004.

[44] Frank Verstraete, Valentin Murg, and J Ignacio Cirac. “Matrix product states, projected entangled pair states, and variational renormalization group methods for quantum spin systems”. In: *Advances in Physics* 57.2 (2008), pp. 143–224.

[45] Daniel C Cabra, Andreas Honecker, and Pierre Pujol. *Modern theories of many-particle systems in condensed matter physics*. Vol. 843. Springer Science & Business Media, 2012.

[46] Román Orús. “A practical introduction to tensor networks: Matrix product states and projected entangled pair states”. In: *Annals of Physics* 349 (Oct. 2014), pp. 117–158. ISSN: 0003-4916. DOI: 10.1016/j.aop.2014.06.013. URL: http://dx.doi.org/10.1016/j.aop.2014.06.013.

[47] Steven R White. “Density matrix formulation for quantum renormalization groups”. In: *Physical review letters* 69.19 (1992), p. 2863.

[48] Steven R White. “Density-matrix algorithms for quantum renormalization groups”. In: *Physical Review B* 48.14 (1993), p. 10345.

[49] Frank Verstraete, Diego Porras, and J Ignacio Cirac. “Density matrix renormalization group and periodic boundary conditions: A quantum information perspective”. In: *Physical review letters* 93.22 (2004), p. 227205.

[50] Andrew John Daley et al. “Time-dependent density-matrix renormalization-group using adaptive effective Hilbert spaces”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2004.04 (2004), P04005.

[51] M. Manjunath et al. “Functional analysis of low-grade glioma genetic variants predicts key target genes and transcription factors”. In: *Neuro Oncol* (Nov. 2020). DOI: 10.1093/neuonc/noaa248.

[52] Y. Zhang et al. “The Cancer-Associated Genetic Variant Rs3903072 Modulates Immune Cells in the Tumor Microenvironment”. In: *Front Genet* 10 (2019), p. 754.

[53] J. Ernst and M. Kellis. "ChromHMM: automating chromatin-state discovery and characterization". In: *Nat Methods* 9.3 (Feb. 2012), pp. 215–216.

[54] M. R. Corces et al. "The chromatin accessibility landscape of primary human cancers". In: *Science* 362.6413 (Oct. 2018).

[55] A. Nott et al. "Brain cell type-specific enhancer-promoter interactome maps and disease-risk association". In: *Science* 366.6469 (Nov. 2019), pp. 1134–1139.

[56] J. K. Pritchard and M. Przeworski. "Linkage disequilibrium in humans: models and data". In: *Am J Hum Genet* 69.1 (July 2001), pp. 1–14.

[57] M. Slatkin. "Linkage disequilibrium–understanding the evolutionary past and mapping the medical future". In: *Nat Rev Genet* 9.6 (June 2008), pp. 477–485.

[58] M. J. Machiela and S. J. Chanock. "LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants". In: *Bioinformatics* 31.21 (Nov. 2015), pp. 3555–3557.

[59] D. J. Brat et al. "Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas". In: *N Engl J Med* 372.26 (June 2015), pp. 2481–2498.

[60] M. Ceccarelli et al. "Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma". In: *Cell* 164.3 (Jan. 2016), pp. 550–563.

[61] S. Das et al. "Next-generation genotype imputation service and methods". In: *Nat Genet* 48.10 (Oct. 2016), pp. 1284–1287.

[62] P. R. Loh et al. "Reference-based phasing using the Haplotype Reference Consortium panel". In: *Nat Genet* 48.11 (Nov. 2016), pp. 1443–1448.

[63] P. R. Loh, P. F. Palamara, and A. L. Price. "Fast and accurate long-range phasing in a UK Biobank cohort". In: *Nat Genet* 48.7 (July 2016), pp. 811–816.

[64] M. V. Rockman and L. Kruglyak. "Genetics of global gene expression". In: *Nat Rev Genet* 7.11 (Nov. 2006), pp. 862–872.

[65] A. C. Nica and E. T. Dermitzakis. "Expression quantitative trait loci: present and future". In: *Philos Trans R Soc Lond B Biol Sci* 368.1620 (2013), p. 20120362.

[66] Y. Gong et al. "Stratification of TAD boundaries reveals preferential insulation of super-enhancers by strong boundaries". In: *Nat Commun* 9.1 (Feb. 2018), p. 542.

[67] Yoav Benjamini and Yosef Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1 (1995), pp. 289–300. ISSN: 00359246. URL: http://www.jstor.org/stable/2346101.

[68] Y. Zhang et al. "Integrative Genomic Analysis Predicts Causative Cis-Regulatory Mechanisms of the Breast Cancer-Associated Genetic Variant rs4415084". In: *Cancer Res* 78.7 (Apr. 2018), pp. 1579–1591.

[69] A. I. Finnegan et al. "Epigenetic engineering of yeast reveals dynamic molecular adaptation to methylation stress and genetic modulators of specific DNMT3 family members". In: *Nucleic Acids Res* 48.8 (May 2020), pp. 4081–4099.

[70] A. Finnegan and J. S. Song. "Maximum entropy methods for extracting the learned features of deep neural networks". In: *PLoS Comput Biol* 13.10 (Oct. 2017), e1005836.

[71] C. E. Grant, T. L. Bailey, and W. S. Noble. "FIMO: scanning for occurrences of a given motif". In: *Bioinformatics* 27.7 (Apr. 2011), pp. 1017–1018.

[72] A. Mathelier et al. "JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles". In: *Nucleic Acids Res* 44.D1 (Jan. 2016), pp. D110–115.

[73] I. V. Kulakovskiy et al. "HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models". In: *Nucleic Acids Res* 44.D1 (Jan. 2016), pp. D116–125.

[74] E. Wingender. "The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation". In: *Brief Bioinform* 9.4 (July 2008), pp. 326–332.

[75] A. Jolma et al. "DNA-binding specificities of human transcription factors". In: *Cell* 152.1-2 (Jan. 2013), pp. 327–339.

[76] H. Li. "A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data". In: *Bioinformatics* 27.21 (Nov. 2011), pp. 2987–2993.

[77] François Chollet et al. *Keras*. https://keras.io. 2015.

[78] Martın Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: https://www.tensorflow.org/.

[79] G. E. Crooks et al. "WebLogo: a sequence logo generator". In: *Genome Res* 14.6 (June 2004), pp. 1188–1190.

[80] E. Bernstein et al. “Role for a bidentate ribonuclease in the initiation step of RNA interference”. In: *Nature* 409.6818 (Jan. 2001), pp. 363–366.

[81] H. Siomi and M. C. Siomi. “On the road to reading the RNA-interference code”. In: *Nature* 457.7228 (Jan. 2009), pp. 396–404.

[82] P. D. Zamore et al. “RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals”. In: *Cell* 101.1 (Mar. 2000), pp. 25–33.

[83] A. Vermeulen et al. “The contributions of dsRNA structure to Dicer specificity and efficiency”. In: *RNA* 11.5 (May 2005), pp. 674–682.

[84] D. Castanotto and J. J. Rossi. “The promises and pitfalls of RNA-interference-based therapeutics”. In: *Nature* 457.7228 (Jan. 2009), pp. 426–433.

[85] P. Ahlquist. “RNA-dependent RNA polymerases, viruses, and RNA silencing”. In: *Science* 296.5571 (May 2002), pp. 1270–1273.

[86] R. Cowper-Sal lari et al. “Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression”. In: *Nat Genet* 44.11 (Nov. 2012), pp. 1191–1198.

[87] S. D. Bailey et al. “ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters”. In: *Nat Commun* 2 (Feb. 2015), p. 6186.

[88] X. Zhang et al. “Integrative functional genomics identifies an enhancer looping to the SOX9 gene disrupted by the 17q24.3 prostate cancer risk locus”. In: *Genome Res* 22.8 (Aug. 2012), pp. 1437–1446.

[89] T. H. Kim and J. Dekker. “ChIP-Quantitative Polymerase Chain Reaction (ChIP-qPCR)”. In: *Cold Spring Harb Protoc* 2018.5 (May 2018).

[90] B. Li et al. “Genotyping with TaqMAMA”. In: *Genomics* 83.2 (Feb. 2004), pp. 311–320.

[91] A. P. Boyle et al. “Annotation of functional variation in personal genomes using RegulomeDB”. In: *Genome Res* 22.9 (Sept. 2012), pp. 1790–1797.

[92] D. Y. Lin et al. “Analysis of the interaction between Zinc finger protein 179 (Znf179) and promyelocytic leukemia zinc finger (Plzf)”. In: *J Biomed Sci* 20 (Dec. 2013), p. 98.

[93] C. L. Hsieh et al. “PLZF, a tumor suppressor genetically lost in metastatic castration-resistant prostate cancer, is a mediator of resistance to androgen deprivation therapy”. In: *Cancer Res* 75.10 (May 2015), pp. 1944–1948.

[94] J. B. Wang et al. “Tumor suppressor PLZF regulated by lncRNA ANRIL suppresses proliferation and epithelial mesenchymal transformation of gastric cancer cells”. In: *Oncol Rep* 41.2 (Feb. 2019), pp. 1007–1018.

[95] H. Shen et al. “PLZF inhibits proliferation and metastasis of gallbladder cancer by regulating IFIT2”. In: *Cell Death Dis* 9.2 (Jan. 2018), p. 71.

[96] Y. Jin, H. Z. Nenseth, and F. Saatcioglu. “Role of PLZF as a tumor suppressor in prostate cancer”. In: *Oncotarget* 8.41 (Sept. 2017), pp. 71317–71324.

[97] A. Khan et al. “JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework”. In: *Nucleic Acids Res* 46.D1 (Jan. 2018), pp. D260–D266.

[98] M. Koubi et al. “Regulation of the positive transcriptional effect of PLZF through a non-canonical EZH2 activity”. In: *Nucleic Acids Res* 46.7 (Apr. 2018), pp. 3339–3350.

[99] F. Felicetti et al. “Role of PLZF in melanoma progression”. In: *Oncogene* 23.26 (June 2004), pp. 4567–4576.

[100] R. M. Hobbs and P. P. Pandolfi. “Shape-shifting and tumor suppression by PLZF”. In: *Oncotarget* 1.1 (May 2010), pp. 3–5.

[101] S. Agrawal Singh et al. “PLZF targets developmental enhancers for activation during osteogenic differentiation of human mesenchymal stem cells”. In: *Elife* 8 (Jan. 2019).

[102] R. Kommagani et al. “The Promyelocytic Leukemia Zinc Finger Transcription Factor Is Critical for Human Endometrial Stromal Cell Decidualization”. In: *PLoS Genet* 12.4 (Apr. 2016), e1005937.

[103] J. T. Robinson et al. “Integrative genomics viewer”. In: *Nat Biotechnol* 29.1 (Jan. 2011), pp. 24–26.

[104] J. E. Eckel-Passow et al. “Using germline variants to estimate glioma and subtype risks”. In: *Neuro Oncol* 21.4 (Mar. 2019), pp. 451–461.

[105] K. Labreche et al. “Diffuse gliomas classified by 1p/19q co-deletion, TERT promoter and IDH mutation status are associated with specific genetic risk loci”. In: *Acta Neuropathol* 135.5 (May 2018), pp. 743–755.

[106] J. MacArthur et al. “The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)”. In: *Nucleic Acids Res* 45.D1 (Jan. 2017), pp. D896–D901.

[107] S. T. Sherry et al. “dbSNP: the NCBI database of genetic variation”. In: *Nucleic Acids Res* 29.1 (Jan. 2001), pp. 308–311.

[108] F. Schmidt et al. "Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction". In: *Nucleic Acids Res* 45.1 (Jan. 2017), pp. 54–66.

[109] A. Nawshad et al. "TGFbeta3 inhibits E-cadherin gene expression in palate medial-edge epithelial cells through a Smad2-Smad4-LEF1 transcription complex". In: *J Cell Sci* 120.Pt 9 (May 2007), pp. 1646–1653.

[110] G. Agrimi et al. "Identification of the human mitochondrial S-adenosylmethionine transporter: bacterial expression, reconstitution, functional characterization and tissue distribution". In: *Biochem J* 379.Pt 1 (Apr. 2004), pp. 183–190.

[111] A. Menga et al. "SLC25A26 overexpression impairs cell function via mtDNA hypermethylation and rewiring of methyl metabolism". In: *FEBS J* 284.6 (Mar. 2017), pp. 967–984.

[112] X. Sun et al. "The degree of mitochondrial DNA methylation in tumor models of glioblastoma and osteosarcoma". In: *Clin Epigenetics* 10.1 (Dec. 2018), p. 157.

[113] Alexander Novikov et al. "Tensor Train Decomposition on TensorFlow (T3F)". In: *Journal of Machine Learning Research* 21.30 (2020), pp. 1–7. URL: http://jmlr.org/papers/v21/18-008.html.

[114] R Hübener, V Nebendahl, and W Dür. "Concatenated tensor network states". In: *New Journal of Physics* 12.2 (Feb. 2010), p. 025004. ISSN: 1367-2630.

[115] Charles F Van Loan. "Tensor network computations in quantum chemistry". In: (2008). URL: http://www.cs.cornell.edu/cv/OtherPdf/ZeuthenCVL.pdf.

[116] Misha Denil et al. "Predicting Parameters in Deep Learning". In: *Advances in Neural Information Processing Systems*. Vol. 26. Curran Associates, Inc., 2013, pp. 2148–2156.

[117] T. N. Sainath et al. "Low-rank matrix factorization for Deep Neural Network training with high-dimensional output targets". In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2013, pp. 6655–6659. DOI: 10.1109/ICASSP.2013.6638949.

[118] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. "Learning representations by back-propagating errors". In: *nature* 323.6088 (1986), pp. 533–536.

[119] Jian Xue, Jinyu Li, and Yifan Gong. "Restructuring of Deep Neural Network Acoustic Models with Singular Value Decomposition". In: *Interspeech*. Jan. 2013.

[120] D. Kardassis et al. "Direct physical interactions between HNF-4 and Sp1 mediate synergistic transactivation of the apolipoprotein CIII promoter". In: *Biochemistry* 41.4 (Jan. 2002), pp. 1217–1228.

[121] R. Lu, E. J. Mucaki, and P. K. Rogan. "Discovery and validation of information theory-based transcription factor and cofactor binding site motifs". In: *Nucleic Acids Res* 45.5 (Mar. 2017), e27.

[122] Michael Dewey. *metap: meta-analysis of significance values*. R package version 1.3. 2020.

[123] Y. Fan et al. "MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data". In: *Genome Biol* 17.1 (Aug. 2016), p. 178.

[124] David Benjamin et al. "Calling Somatic SNVs and Indels with Mutect2". In: *bioRxiv* (2019). DOI: 10.1101/861054.

[125] D. E. Larson et al. "SomaticSniper: identification of somatic point mutations in whole genome sequencing data". In: *Bioinformatics* 28.3 (Feb. 2012), pp. 311–317.

[126] D. C. Koboldt et al. "VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing". In: *Genome Res* 22.3 (Mar. 2012), pp. 568–576.

[127] V. Gleize et al. "CIC inactivating mutations identify aggressive subset of 1p19q codeleted gliomas". In: *Ann Neurol* 78.3 (Sept. 2015), pp. 355–374.

Appendix A

Supplementary Material for Chapter 2

A.1 LGG GWAS SNPs and SNPs in high linkage disequilibrium

We obtained a list of GWAS SNPs from Melin *et al.* [12], passing the combined meta-analysis (eight studies) P-value cutoff of 5×10^{-8} for non-glioblastoma gliomas. This yields 25 GWAS SNPs significantly associated with LGG (Table A.1). 8 SNPs out of these 25 were also found to be significant ($P < 5 \times 10^{-8}$) in glioblastoma. The median odds ratio for the 25 GWAS SNPs was 1.2, where 23 of the 25 SNPs had odds ratio less than 1.5, typical of low-penetrance genetic variants [12]. We then used LDlink [58] to obtain all SNPs in high linkage disequilibrium ($r^2 \geqslant 0.8$, 1000 Genomes Phase 3, EUR population) with the 25 GWAS SNPs and analyzed the functional footprinting of 280 SNPs in total.

GWAS SNP ID	Reference Allele	Alternative Allele	Risk Allele
rs10069690	C	T	T
rs10131032	G	A	G
rs11196067	A	T	A
rs11598018	C	A	C
rs11599775	G	A	G
rs11706832	A	C	C
rs12076373	G	C	G
rs1275600	T	A	T
rs12803321	G	C	G
rs1801591	G	A	A
rs2297440	T	C	C
rs2736100	C	A	C
rs3751667	C	T	T
rs4252707	G	A	A
rs4977756	G	A	G
rs498872	A	G	A
rs55705857	A	G	G
rs6010620	A	G	G
rs634537	T	G	G
rs648044	A	G	A
rs7107785	T	C	T
rs75061358	T	G	G
rs7572263	A	G	A
rs77633900	G	C	C
rs78378222	T	G	G

Table A.1: 25 GWAS SNPs significantly associated with LGG

A.2 Motif permutation test

The motif permutation test described in Appendix A.2 was developed by Dr. Yi Zhang and Dr. Mohith Manjunath [68]. The originally published method [68] is rephrased below.

Given a TF's position probability matrix (PPM) of length L, the information content (IC) at position i ($i \in \{1, 2, \ldots, L\}$) could be written as:

$$IC(i) = \sum_{j \in \{A,C,G,T\}} p_{ij} log \frac{p_{ij}}{q_j}, \tag{A.1}$$

where p_{ij} denotes the probability of nucleotide j ($j \in \{A, C, G, T\}$) at position i, and q_j denotes the background frequency of nucleotide j. We used $q_A = 0.3$, $q_C = 0.2$, $q_G = 0.2$ and $q_T = 0.3$ throughout the motif permutation test. For a genomic sequence $\boldsymbol{S} = s_1 s_2 \ldots s_L$ (s_i denotes the nucleotide at position i of the sequence) of length L, the score for the match of the sequence to a motif (length L) is defines as:

$$score(\boldsymbol{S}) = \sum_{i=1}^{L} log \frac{p_{is_i}}{q_{s_i}}. \tag{A.2}$$

Thus, if we assume the SNP at position k, and use $\boldsymbol{S} = s_1 s_2 \ldots s_k \ldots s_L$ to denote the motif-matching sequence detected by FIMO, while using $\boldsymbol{S'} = s_1 s_2 \ldots s'_k \ldots s_L$ to denote the sequence containing an alternative allele of the SNP, the score difference can be written as:

$$D(\boldsymbol{S}, \boldsymbol{S'}) = score(\boldsymbol{S}) - score(\boldsymbol{S'}) = log \frac{p_{ks_k} q_{s'_k}}{p_{ks'_k} q_{s_k}}. \tag{A.3}$$

We then generated a null distribution of score change by introducing n ($n = 5000$) random single nucleotide mutations at position i of sequence $\boldsymbol{S}$, converting s_i to a random nucleotide s_i^a, where $s_i^a \in \{A, C, G, T\}$, $i \in \{1, 2, \ldots, L\}$. In order to simulate neutral mutations, random mutations are introduced in two steps. First, the choice of mutation position i in the sequence is determined by sampling from a distribution inversely proportional to the position information content:

$$P_{m_1}(i) = \frac{1}{IC(i)} \cdot \frac{1}{\sum_{i=1}^{L} \frac{1}{IC(i)}}. \tag{A.4}$$

Here, $P_{m_1}(i)$ denotes the probability of a neural mutation being introduced at position i, and is normalized by a constant $\sum_{i=1}^{L} \frac{1}{IC(i)}$. Second, given the mutation position i, the probability of converting s_i to nucleotide

s_i^a ($s_i^a \in \{A, C, G, T\}$), $P_{m_2}(s_i^a|i)$, is proportional to $p_{is_i^a}$ in the position probability matrix:

$$P_{m_2}(s_i^a|i) = p_{is_i^a}. \tag{A.5}$$

Since the probability of converting s_i to s_i^a is equal to 0 when $p_{is_i^a} = 0$, instead of using the original position probability matrix, we added a pseudocount of 5% to p_{ij} ($i \in \{1, 2, \ldots, L\}$, $j \in \{A, C, G, T\}$), and normalized the sum of probability at each position to 1. The neutral mutation probability is thus written as:

$$P_m(s_i^a) = P_{m_1}(i) \cdot P_{m_2}(s_i^a|i). \tag{A.6}$$

After performing the above steps, we now obtain the new sequence $\boldsymbol{S_t^a}$ for the t^{th} simulation ($t \in \{1, 2, \ldots, n\}$). We could then compute the score difference $D(\boldsymbol{S}, \boldsymbol{S_t^a})$ for each mutated sequence $\boldsymbol{S_t^a}$. The P-value of the motif disruption by the single nucleotide change of the SNP is thus given by:

$$P_{value}(\boldsymbol{S}, \boldsymbol{S'}) = \frac{\sum_{t=1}^{n} 1\{|D(\boldsymbol{S}, \boldsymbol{S_t^a}) > D(\boldsymbol{S}, \boldsymbol{S'})|\}}{n}. \tag{A.7}$$

We repeated the calculation of P-value for 100 times, and reported the average P-value obtained. Setting permutation test P-value threshold as 0.05, we thus selected all TFs whose binding affinity were significantly perturbed by the SNP.

Appendix B

Supplementary Material for Chapter 3

B.1 Allele-specific ATAC-seq read counts of TCGA LGG samples

For the ATAC-seq data, there were 13 samples in total with the aligned reads (hg38) in BAM format. We extracted the read counts by allele at the rs648044 location using bcftools mpileup [76] option. We considered only the bases with a Phred quality score of at least 20. Out of 13 samples, three were removed because the imputed genotype status of rs648044 was not heterozygous. The genotype status of rs648044 in one of the samples was imputed to be homozygous but was retained because the ATAC-seq reads showed high coverage for both alleles at the SNP location. For each sample, the significance of the skew between the two alleles was evaluated using a binomial test. The resulting P-values were then combined using the Fisher's method from the R package metap [122] (Table B.1).

Aliquot barcode	Allele A	Allele G	Binomial P-value
TCGA-P5-A735-01A-31-A617-42-X017-S03	47	45	4.59E-01
TCGA-E1-A7YI-01A-31-A617-42-X020-S08	84	83	5.00E-01
TCGA-DU-5870-02A-21-A646-42-X030-S02	61	54	2.88E-01
TCGA-FG-A4MU-01B-31-A615-42-X016-S10	61	40	2.30E-02
TCGA-P5-A72W-01A-31-A617-42-X018-S01	69	40	3.52E-03
TCGA-DU-6407-02B-21-A645-42-X036-S01	112	98	1.85E-01
TCGA-P5-A72X-01A-31-A617-42-X014-S09	8	5	2.91E-01
TCGA-FG-A4MY-01A-31-A616-42-X015-S02	25	16	1.06E-01
TCGA-F6-A8O3-01A-31-A617-42-X013-S07	35	33	4.52E-01
TCGA-W9-A837-01A-31-A617-42-X019-S03	16	15	5.00E-01
Total	518	429	1.00E-02 (Fisher's method)

Table B.1: Allele-specific ATAC-seq read counts covering the GWAS SNP rs648044 and the corresponding two-sided binomial test P-values of TCGA LGG samples. The P-values were combined using the Fisher's method.

B.2 Electrophoretic Mobility Shift Assay (EMSA)

EMSA was performed with the mixture of the recombinant MafF protein and four different DNA oligonucleotides: Positive Control (PC), rs648044 locus containing the A allele, rs648044 locus containing the G allele, Negative Control (NC). The sequences of the oligonucleotides and recombinant MafF are given below.

Two complementary oligonucleotide strands were mixed in a PCR tube and hybridized by increasing the temperature to 98°C and lowering by 5°C every 5 minutes until the temperature reached 4°C using a thermocycler (Mastercycler Personal, Eppendorf). For the binding of MafF, 4 pmole of the hybridized oligonucleotide and 32 pmole of MafF were mixed with 10 mM $MgCl_2$ in T50 buffer (10 mM Tris-Cl, 50 mM NaCl, pH 8.0). For MafF negative samples, the same materials were mixed, except for MafF. The resulting mixtures were incubated at 37°C for 30 minutes. After the incubation, the mixtures were subjected to polyacrylamide gel electrophoresis, fluorescently labeled by SYBR Gold Nucleic Acid Gel Stain (S11494, ThermoFisher) and visualized on a UV illuminator (Dyna Light UV Transilluminator, Labnet International, Inc.). The recombinant MafF is purchased from Abnova (GST-tag removed, catalog number: H00023764-Q01), and the recombinant MafF sequence is: MSVDPLSSKALKIKRELSENTPHLSDEALMGLSVRELNRHLRGLSAEEVTRLKQRRRTLKNRGYAASCRVKRVCQKEELQKQKSELEREVDKLARENAAMRLELDALRGK. The oligonucleotide sequences representing the rs648044 locus are given below, where the core binding motif of MAFF is highlighted and underlined:

81 bp flanking sequence harboring the rs648044-A allele:

5'-CCTTGCACTGGCACATTCCTGCTGTTTTCTTCTGC**TCAGCA**GAGCCGAACGGCTCTCACTTCCTGGCTAGCTCTGTGTGCT-3'

81 bp flanking sequence harboring the rs648044-G allele:

5'-CCTTGCACTGGCACATTCCTGCTGTTTTCTTCTGC**TCAGCG**GAGCCGAACGGCTCTCACTTCCTGGCTAGCTCTGTGTGCT-3'

The control sequences were designed based on ChIP-seq results (Appendix B.3). All oligonucleotides were purchased from Integrated DNA Technologies.

B.3 EMSA experiment MAFF positive control (PC) and negative control (NC) sequences

We first obtained MAFF ChIP-seq peaks in HepG2, K562 and HelaS3 cell lines from ENCODE (HepG2: ENCFF611VKE; K562: ENCFF864KPF; HelaS3: ENCFF575MOS). We then ranked the peaks by q-value in each cell type to obtain top peak regions. We intersected the top peak regions and scanned the sequences using FIMO [71], keeping only the consensus peaks containing a MAFF binding motif. Upon visual inspection of raw ChIP-seq signals of the top remaining candidates, we chose chr14:77423081-77423161 (hg19) as

the 81bp-long positive control sequence centered around a MAFF core binding motif (TCAGCA). For the negative control sequence, we first scanned the 81 bp flanking sequence harboring the rs648044-A allele and obtained all sub-sequences which might partially contain a core MAFF motif. We then randomly permuted the nucleotides in those subsequences. We checked that the resulting negative control sequence satisfied the following two criteria: (1) there were no more than three adjacent nucleotides of MAFF core binding motif (TCAGCA or its reverse compliment TGCTGA); (2) the GC content of the negative control sequence was approximately the same as the original sequence harboring the rs648044-A allele. The positive control and negative control sequences are provided below, where the core binding motif of MAFF is highlighted and underlined in the positive control:

Positive control sequence:

5'-GTTCCCGCCGCCCGGAGGCTCATTTGTCACCCGC**TGCTGA**C**TCAGCA**CTTCTGCAGAAG GCTTTTCCCTCCGCTTTGGAGG-3'

Negative control sequence:

5'-ATAACACAGAGCTAGCCACGAAGTGAGAGCCGTTCGCCTCGTGGACGGAGAAGAAAACACC CGGAATGTGCCAGTGCAATT-3'

B.4 Cell Culture, RNAi and RNA expression

The cell line SF10417 (from UCSF) derived from a human *IDH1*^{R132H} mutant, *TERT* promoter-mutant, 1p/19q-codeleted oligodendroglioma was used to assess the effect of MAFF knockdown on *ZBTB16* and *NCAM1* expression. The cells were grown in NeuroCult NS-A media (STEMCELL Technologies) supplemented with L-Glutamine, B27, N2, Sodium Pyruvate and Pen/Strep (Life Technologies) in the presence of growth factors bFGF, EGF (STEMCELL Technologies), and PDGFAA (PeproTech). Lentivirus were produced with short-hairpin RNA (shRNA) with either a non-target shRNA control vector or a vector designed to reduce the expression of *MAFF* (n = 3 independent constructs, Sigma Mission). Cells were infected with a MOI of 1, followed by selection of transduced cells with puromycin. Populations were subcloned and total RNA was isolated for three independent clones for each of the vectors (Qiagen). First strand cDNA synthesis was completed with SuperScript IV VILO (Invitrogen) and gene expression was measured with TaqMan probes according to manufacture guidelines for genes *18S*, *MAFF*, *ZBTB16* and *NCAM1* (Applied Biosystems).

B.5 The effect of MAFF RNAi knockdown on *NCAM1* expression

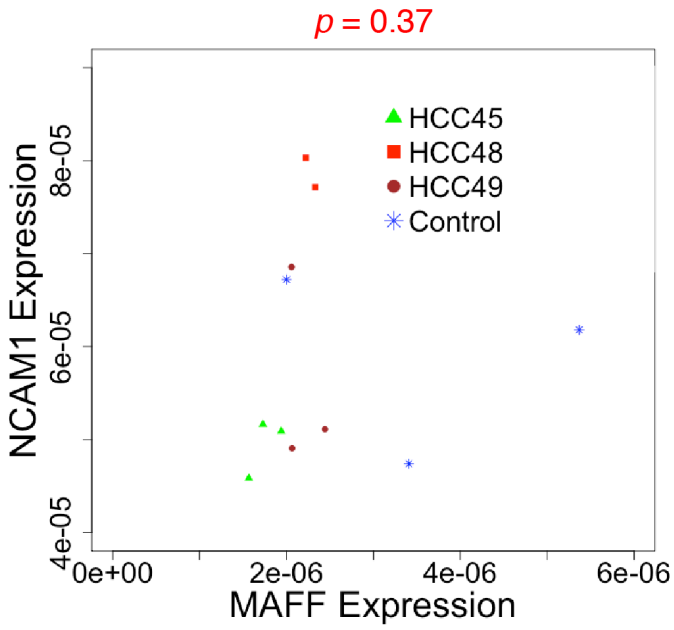


Figure B.1: The results of MAFF RNAi knockdown experiment showing insignificant effect on *NCAM1* expression. One-sided t-test P-value between the control group and the combined group of three independent shRNA clones is shown on top of the figure.

B.6 *CIC* inactivating mutations

We acquired the mutation calls from the 4 somatic variant calling pipelines: MuSE [123], MuTect2 [124], SomaticSniper [125] and VarScan2 [126], available in the GDC Data Portal [30]. For each of the 4 pipelines, we then obtained all *CIC* inactivating mutations based on whether the PolyPhen column contained the "probably_damaging" term or the IMPACT column was "HIGH" [127]. To reduce false positives, we retained only those inactivating mutation calls detected by at least 2 of the 4 pipelines.

B.7 ZBTB16 ChIP-seq data from Gene Expression Omnibus

ZBTB16 (PLZF) ChIP-seq datasets were obtained from Gene Expression Omnibus accession numbers GSE125166 (human mesenchymal stem cells), GSE75115 (human endometrial stromal cells) and GSE109619 (acute myelogenous leukemia cell line KG1).

Appendix C

Supplementary Material for Chapter 4

C.1 GWAS SNP rs12803321 and its high LD SNPs

RS number	Chromosome	Position (hg19)	Alleles	r^2	GWAS SNP correlated alleles
rs12803321	11	118480115	G/C	1	-
rs67307131	11	118480223	T/C	0.98	G=T, C=C
rs12225399	11	118480285	G/C	0.97	G=G, C=C
rs7125115	11	118478330	G/A	0.90	G=G, C=A

Table C.1: GWAS SNP rs12803321 and its three high LD ($r^2 \geqslant 0.8$, 1000 Genomes Phase 3, EUR) SNPs.

C.2 Phased allele-specific expression of *PHLDB1*

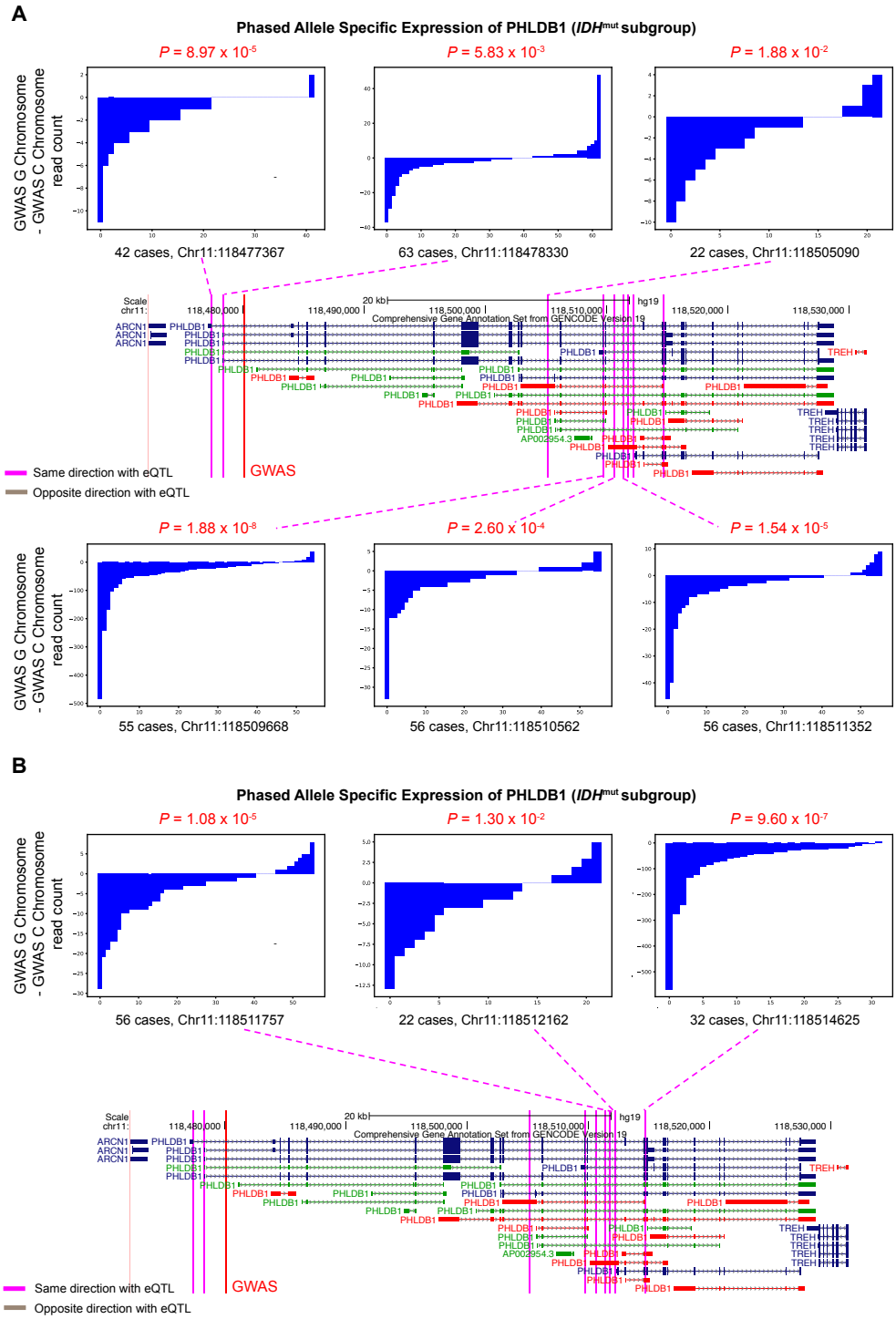


Figure C.1: (A, B) Phased allele-specific transcription pattern of *PHLDB1* at 9 exonic SNPs in the TCGA-LGG "IDH^{mut} only" subgroup. All 9 SNPs show higher transcription emanating from the rs12803321-C haplotype, consistent with the eQTL analysis. For each exonic SNP, RNA-Seq read count differences between the two chromosomes harboring the rs12803321-G allele and the rs12803321-C allele are sorted across patients and shown as bar plots. The P-value from the Wilcoxon signed-rank sum test is shown at the top of each bar plot. The genomic locations of these 9 exonic SNPs are shown in the GENCODE Version 19 track.

C.3 Candidate TFs perturbed by rs7125115

motif from	motif to	motif id	motif source	TF	strand	hit allele (G/A)	Fimo *P*-value	permut *P*-value
118478329	118478346	MA0729.1	JASPAR	RARA	-	A	0.000995	0.04
118478320	118478335	MA0868.1	JASPAR	SOX8	+	A	0.000656	0.02

Table C.2: Motifs of candidate TFs perturbed by rs7125115. Columns from left to right: motif start coordinate in chr11 (hg19); motif end coordinate in chr11 (hg19); motif id from one of the four databases - JASPAR [72], HOCOMOCOv10 [73], TRANSFAC [74] and Jolma2013 [75]; motif source, one of the four databases - JASPAR [72], HOCOMOCOv10 [73], TRANSFAC [74] and Jolma2013 [75]; transcription factor name; strand harboring the motif; rs7125115 allele harbored by the motif; *P*-value from Fimo output; *P*-value from the permutation test.

Group	TF	TF avg expr	GG number	GA number	AA number	r_{GG}	r_{GA}	r_{AA}
All TCGA-LGG samples	RARA	9.72	264	172	42	0.23	0.12	0.21
IDH^{mut} only and TP	RARA	9.77	195	105	22	0.25	0.30	0.42
IDH^{mut} only	RARA	9.68	118	65	10	0.11	0.32	0.34
TP	RARA	9.91	77	40	12	0.32	0.24	0.63
All TCGA-LGG samples	SOX8	13.49	264	172	42	0.24	0.39	0.43
IDH^{mut} only and TP	SOX8	13.96	195	105	22	0.33	0.49	0.27
IDH^{mut} only	SOX8	13.76	118	65	10	0.31	0.59	0.25
TP	SOX8	14.27	77	40	12	0.25	0.34	0.66

Table C.3: The Pearson's correlation coefficient (r) between TF and *PHLDB1* stratified into rs7125115 GG, GA and AA genotype groups. Columns from left to right: patient group; transcription factor name; TF average expression, defined as $\overline{log_2(\text{RSEM}+1)}$; patient number in GG genotype group; patient number in GA genotype group; patient number in AA genotype group; Pearson's correlation coefficient in GG genotype group; Pearson's correlation coefficient in GA genotype group; Pearson's correlation coefficient in AA genotype group.

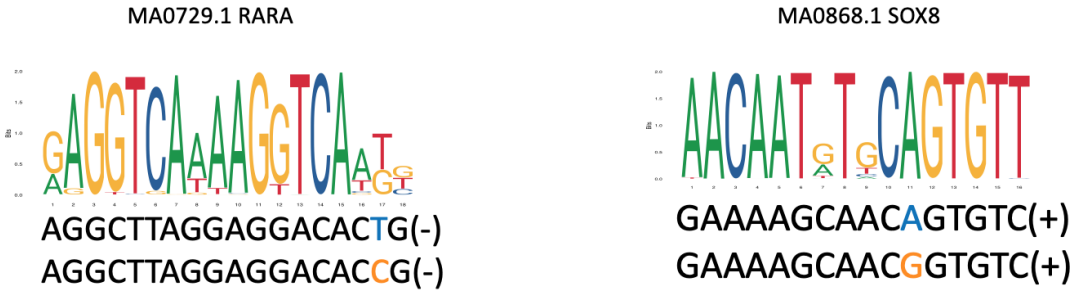


Figure C.2: Motif logo of each candidate TF listed in Table C.2 and variants of the flanking sequences harboring rs7125115 risk allele G (C for - strand) or rs7125115 non-risk allele A (T for - strand). TFs are predicted to have higher binding affinity to the sequences on top than the sequences at bottom.

C.4 Candidate TFs perturbed by rs12803321

motif from	motif to	motif id	motif source	TF	strand	hit allele (G/C)	Fimo *P*-value	permut *P*-value
118480112	118480122	FOSB_HUMAN H10MO.C	HOCOMOCO	FOSB	+	G	0.000142	0.013
118480111	118480121	SUH_HUMAN H10MO.C	HOCOMOCO	RBPJ	-	C	0.000436	0.0009
118480109	118480118	RXRB_HUMAN H10MO.C	HOCOMOCO	RXRB	+	C	0.000452	0.034
118480107	118480128	SP2_HUMAN H10MO.C	HOCOMOCO	SP2	-	C	0.000382	0.014
118480109	118480118	TFE3_HUMAN H10MO.C	HOCOMOCO	TFE3	-	G	0.000978	0.014
118480106	118480125	P53_HUMAN H10MO.B	HOCOMOCO	TP53	-	C	0.000554	0.035

Table C.4: Motifs of candidate TFs perturbed by rs12803321. Columns from left to right: motif start coordinate in chr11 (hg19); motif end coordinate in chr11 (hg19); motif id from one of the four databases - JASPAR [72], HOCOMOCOv10 [73], TRANSFAC [74] and Jolma2013 [75]; motif source, one of the four databases - JASPAR [72], HOCOMOCOv10 [73], TRANSFAC [74] and Jolma2013 [75]; transcription factor name; strand harboring the motif; rs12803321 allele harbored by the motif; *P*-value from Fimo output; *P*-value from the permutation test.

Group	TF	TF avg expr	GG number	GC number	CC number	r_{GG}	r_{GC}	r_{CC}
All TCGA-LGG samples	FOSB	7.55	273	172	29	-0.23	-0.11	-0.06
IDH^{mut} only and TP	FOSB	7.46	200	106	17	-0.18	-0.12	-0.08
IDH^{mut} only	FOSB	7.40	120	65	8	-0.175	-0.172	0.048
TP	FOSB	7.54	80	41	9	-0.18	-0.13	-0.37
All TCGA-LGG samples	RBPJ	9.86	273	172	29	0.25	0.28	0.47
IDH^{mut} only and TP	RBPJ	10.00	200	106	17	0.25	0.21	0.31
IDH^{mut} only	RBPJ	10.03	120	65	8	0.29	0.29	0.36
TP	RBPJ	9.94	80	41	9	0.21	0.25	-0.07
All TCGA-LGG samples	RXRB	10.54	273	172	29	0.14	0.25	0.25
IDH^{mut} only and TP	RXRB	10.63	200	106	17	0.24	0.28	0.16
IDH^{mut} only	RXRB	10.58	120	65	8	0.14	0.10	0.29
TP	RXRB	10.71	80	41	9	0.25	0.41	0.37
All TCGA-LGG samples	SP2	9.54	273	172	29	0.24	0.22	0.47
IDH^{mut} only and TP	SP2	9.55	200	106	17	0.25	0.34	0.47
IDH^{mut} only	SP2	9.59	120	65	8	0.24	0.28	0.36
TP	SP2	9.51	80	41	9	0.35	0.53	0.28
All TCGA-LGG samples	TFE3	11.05	273	172	29	-0.18	-0.08	-0.12
IDH^{mut} only and TP	TFE3	11.03	200	106	17	-0.14	-0.11	-0.03
IDH^{mut} only	TFE3	11.11	120	65	8	-0.11	-0.05	0.14
TP	TFE3	10.92	80	41	9	-0.09	-0.16	-0.07
All TCGA-LGG samples	TP53	9.97	273	172	29	0.26	0.08	0.21
IDH^{mut} only and TP	TP53	10.00	200	106	17	0.25	0.17	0.42
IDH^{mut} only	TP53	9.85	120	65	8	0.27	0.13	0.57
TP	TP53	10.22	80	41	9	0.11	0.22	0.35

Table C.5: The Pearson's correlation coefficient (r) between TF and *PHLDB1* stratified into rs12803321 GG, GC and CC genotype groups. Columns from left to right: patient group; transcription factor name; TF average expression, defined as $\overline{log_2(\text{RSEM} + 1)}$; patient number in GG genotype group; patient number in GC genotype group; patient number in CC genotype group; Pearson's correlation coefficient in GG genotype group; Pearson's correlation coefficient in GC genotype group; Pearson's correlation coefficient in CC genotype group.

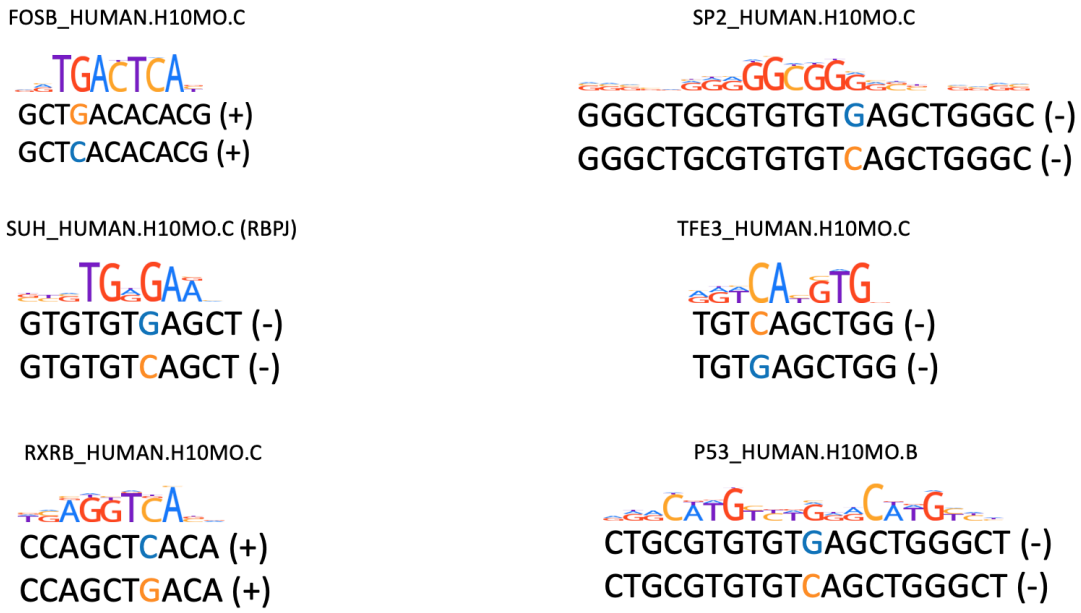


Figure C.3: Motif logo of each candidate TF listed in Table C.4 and variants of the flanking sequences harboring rs12803321 risk allele G (C for - strand) or rs12803321 non-risk allele C (G for - strand). TFs are predicted to have higher binding affinity to the sequences on top than the sequences at bottom.

C.5 Candidate TFs perturbed by rs67307131

motif from	motif to	motif id	motif source	TF	strand	hit allele (T/C)	Fimo P-value	permut P-value
118480218	118480228	M00778 .AhR .matrix	TRANSF-AC Human	AHR	-	C	0.000233	0.0061
118480220	118480227	MA0259.1	JASPAR	ARNT ::HIF1A	-	C	0.00094	0.017
118480212	118480230	M00237 .AhRArnt .matrix	TRANSF-AC Human	AHR ::ARNT	-	C	0.000921	0.0086
118480214	118480232	M00237 .AhRArnt .matrix	TRANSF-AC Human	AHR ::ARNT	-	C	9.20E-05	0.013
118480215	118480228	EGR1_ DBD	jolma2013	EGR1	+	C	0.00046	0.00056
118480215	118480228	EGR1_full	jolma2013	EGR1	+	C	0.000629	0.0038
118480216	118480226	EGR1_full	HOCO-MOCO	EGR1	-	C	0.000591	0.0053
118480217	118480226	MA0825.1	JASPAR	MNT	+	T	0.000758	0.0016
118480215	118480225	MESP1_ HUMAN .H10MO.D	HOCO-MOCO	MESP1	-	T	0.00037	0.014
118480217	118480226	MESP1_ DBD	jolma2013	MESP1	-	T	0.000846	0.03
118480216	118480233	NRF1_ HUMAN .H10MO.A	HOCO-MOCO	NRF1	-	C	0.000714	0.004
118480215	118480225	MA0506.1	JASPAR	NRF1	-	C	0.000655	0.032
118480214	118480223	MA0802.1	JASPAR	TBR1	-	T	0.000293	0.063
118480217	118480225	BHE40_ HUMAN .H10MO.A	HOCO-MOCO	BHLHE40	-	T	0.000792	0.013
118480213	118480226	GMEB2_ DBD_2	jolma2013	GMEB2	+	C	0.000454	0.0043
118480219	118480237	SP1_ HUMAN .H10MO.C	HOCO-MOCO	SP1	-	C	0.000163	0.017
118480221	118480233	M00196 .Sp1 .matrix	TRANSF-AC Human	SP1	-	C	0.000269	0.081
118480213	118480229	KLF15_ HUMAN .H10MO.D	HOCO-MOCO	KLF15	-	C	0.000667	0.01

Table C.6: Motifs of candidate TFs perturbed by rs67307131. Columns from left to right: motif start coordinate in chr11 (hg19); motif end coordinate in chr11 (hg19); motif id from one of the four databases - JASPAR [72], HOCOMOCOv10 [73], TRANSFAC [74] and Jolma2013 [75]; motif source, one of the four databases - JASPAR [72], HOCOMOCOv10 [73], TRANSFAC [74] and Jolma2013 [75]; transcription factor name; strand harboring the motif; rs67307131 allele harbored by the motif; P-value from Fimo output; P-value from the permutation test.

Group	TF	TF avg expr	TT number	TC number	CC number	r_{TT}	r_{TC}	r_{CC}
All TCGA-LGG samples	AHR	8.00	277	167	30	-0.03	-0.13	-0.39
IDH^{mut} only and TP	AHR	7.89	202	102	17	-0.07	-0.23	-0.15
IDH^{mut} only	AHR	8.13	122	62	8	-0.02	-0.17	-0.20
TP	AHR	7.5	80	40	9	0.019	-0.26	-0.30
All TCGA-LGG samples	ARNT	9.55	277	167	30	-0.07	0.09	0.12
IDH^{mut} only and TP	ARNT	9.57	202	102	17	0.098	0.13	0.046
IDH^{mut} only	ARNT	9.58	122	62	8	0.014	0.47	0.043
TP	ARNT	9.55	80	40	9	0.31	-0.53	0.46
All TCGA-LGG samples	HNF1A	12.07	277	167	30	-0.078	-0.12	0.04
IDH^{mut} only and TP	HNF1A	12.09	202	102	17	-0.008	-0.03	-0.16
IDH^{mut} only	HNF1A	12.13	122	62	8	0.04	0.018	-0.29
TP	HNF1A	12.02	80	40	9	-0.04	-0.055	-0.20
All TCGA-LGG samples	ARNT2	12.91	277	167	30	-0.25	0.07	0.00
IDH^{mut} only and TP	ARNT2	12.92	202	102	17	-0.17	-0.009	-0.45
IDH^{mut} only	ARNT2	12.87	122	62	8	-0.24	0.0055	-0.56
TP	ARNT2	13.01	80	40	9	-0.11	-0.16	-0.010
All TCGA-LGG samples	EGR1	11.03	277	167	30	-0.15	-0.03	-0.064
IDH^{mut} only and TP	EGR1	10.78	202	102	17	-0.15	0.043	0.11
IDH^{mut} only	EGR1	11.11	122	62	8	-0.004	0.14	0.52
TP	EGR1	10.30	80	40	9	-0.225	-0.122	-0.57
All TCGA-LGG samples	MNT	9.44	277	167	30	-0.09	0.09	-0.17
IDH^{mut} only and TP	MNT	9.48	202	102	17	-0.06	0.11	-0.26
IDH^{mut} only	MNT	9.48	122	62	8	-0.16	0.13	-0.17
TP	MNT	9.48	80	40	9	0.08	0.09	-0.42
All TCGA-LGG samples	MESP1	6.19	277	167	30	-0.03	0.04	-0.03
IDH^{mut} only and TP	MESP1	6.22	202	102	17	-0.09	0.009	0.24
IDH^{mut} only	MESP1	6.11	122	62	8	-0.17	0.07	0.45
TP	MESP1	6.38	80	40	9	-0.07	-0.17	0.35

Continue in the next page

Group	TF	TF avg expr	TT number	TC number	CC number	r_{TT}	r_{TC}	r_{CC}
All TCGA-LGG samples	NRF1	8.77	277	167	30	0.11	0.091	-0.16
IDH^{mut} only and TP	NRF1	8.78	202	102	17	0.14	0.28	-0.04
IDH^{mut} only	NRF1	8.84	122	62	8	0.15	0.36	-0.39
TP	NRF1	8.69	80	40	9	0.29	0.36	-0.28
All TCGA-LGG samples	TBR1	5.07	277	167	30	-0.18	-0.15	-0.26
IDH^{mut} only and TP	TBR1	5.07	202	102	17	-0.18	-0.15	-0.36
IDH^{mut} only	TBR1	4.80	122	62	8	-0.13	-0.18	-0.67
TP	TBR1	5.46	80	40	9	-0.31	-0.24	0.008
All TCGA-LGG samples	BHLHE40	9.85	277	167	30	-0.33	-0.33	-0.45
IDH^{mut} only and TP	BHLHE40	9.60	202	102	17	-0.35	-0.37	-0.21
IDH^{mut} only	BHLHE40	9.65	122	62	8	-0.40	-0.39	-0.41
TP	BHLHE40	9.54	80	40	9	-0.24	-0.33	0.05
All TCGA-LGG samples	GMEB2	8.82	277	167	30	0.09	0.1	-0.20
IDH^{mut} only and TP	GMEB2	8.82	202	102	17	0.13	0.25	0.14
IDH^{mut} only	GMEB2	8.74	122	62	8	-0.09	0.08	0.16
TP	GMEB2	8.95	80	40	9	0.21	0.43	0.19
All TCGA-LGG samples	SP1	10.38	277	167	30	0.20	0.098	0.17
IDH^{mut} only and TP	SP1	10.38	202	102	17	0.22	0.20	0.21
IDH^{mut} only	SP1	10.43	122	62	8	0.31	0.17	-0.34
TP	SP1	10.32	80	40	9	0.13	0.39	0.18
All TCGA-LGG samples	KLF15	10.33	277	167	30	0.055	0.23	0.095
IDH^{mut} only and TP	KLF15	10.38	202	102	17	0.14	0.19	-0.15
IDH^{mut} only	KLF15	10.22	122	62	8	0.086	0.27	0.06
TP	KLF15	10.60	80	40	9	0.11	0.007	-0.014

Table C.7: The Pearson's correlation coefficient (r) between TF and *PHLDB1* stratified into rs67307131 TT, TC and CC genotype groups. Columns from left to right: patient group; transcription factor name; TF average expression, defined as $\overline{log_2(\text{RSEM} + 1)}$; patient number in TT genotype group; patient number in TC genotype group; patient number in CC genotype group; Pearson's correlation coefficient in TT genotype group; Pearson's correlation coefficient in TC genotype group; Pearson's correlation coefficient in CC genotype group.

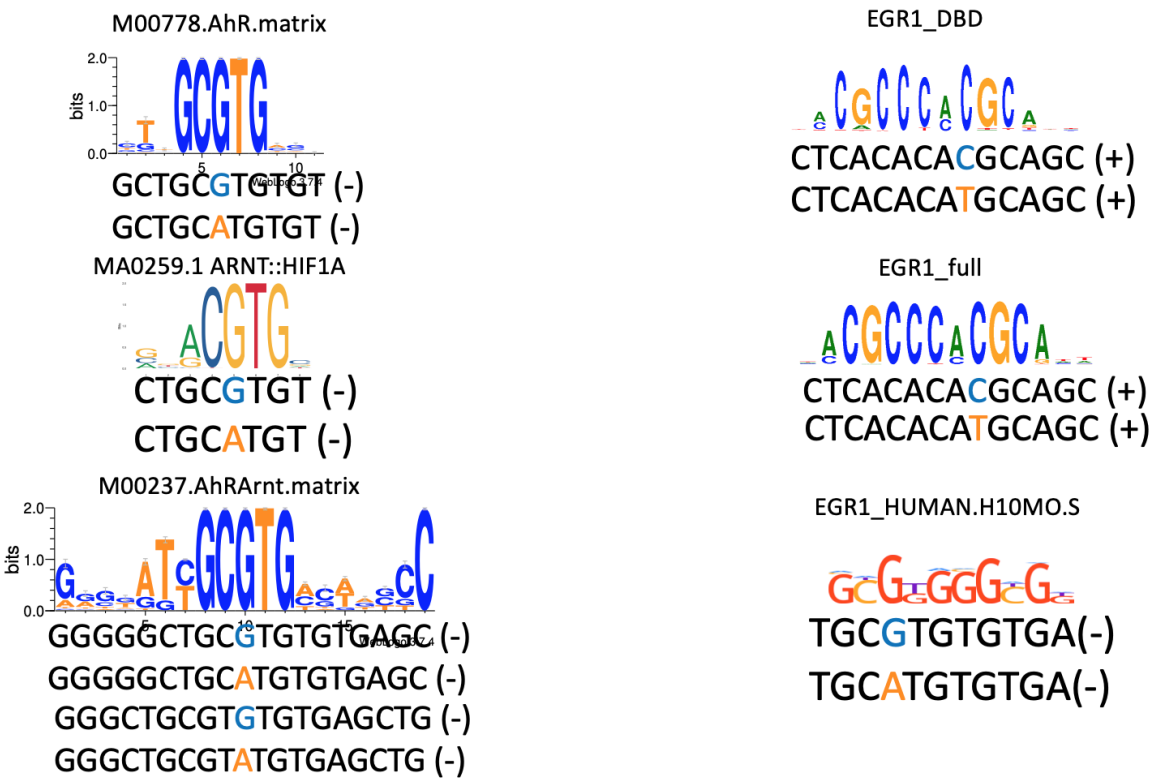


Figure C.4: Motif logo of each candidate TF listed in Table C.6 and variants of the flanking sequences harboring rs67307131 risk allele T (A for - strand) or rs67307131 non-risk allele C (G for - strand). TFs are predicted to have higher binding affinity to the sequences on top than the sequences at bottom.

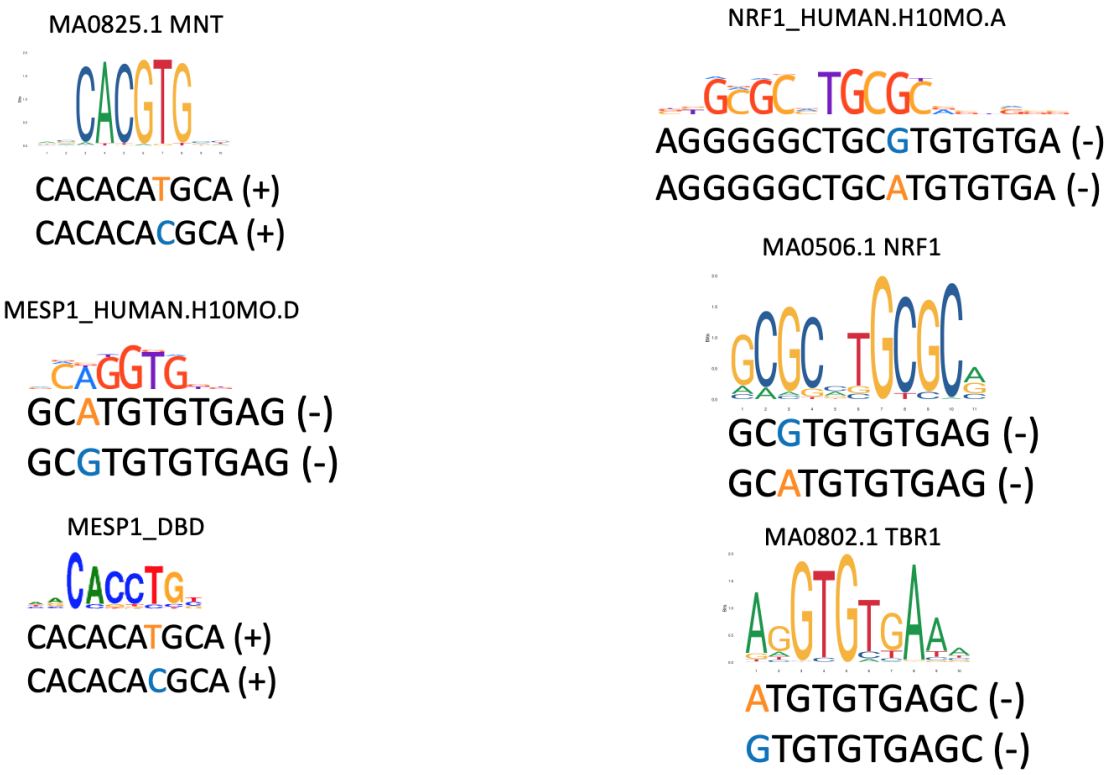


Figure C.5: (Figure C.4 continued) - motif logo of each candidate TF listed in Table C.6 and variants of the flanking sequences harboring rs67307131 risk allele T (A for - strand) or rs67307131 non-risk allele C (G for - strand). TFs are predicted to have higher binding affinity to the sequences on top than the sequences at bottom.

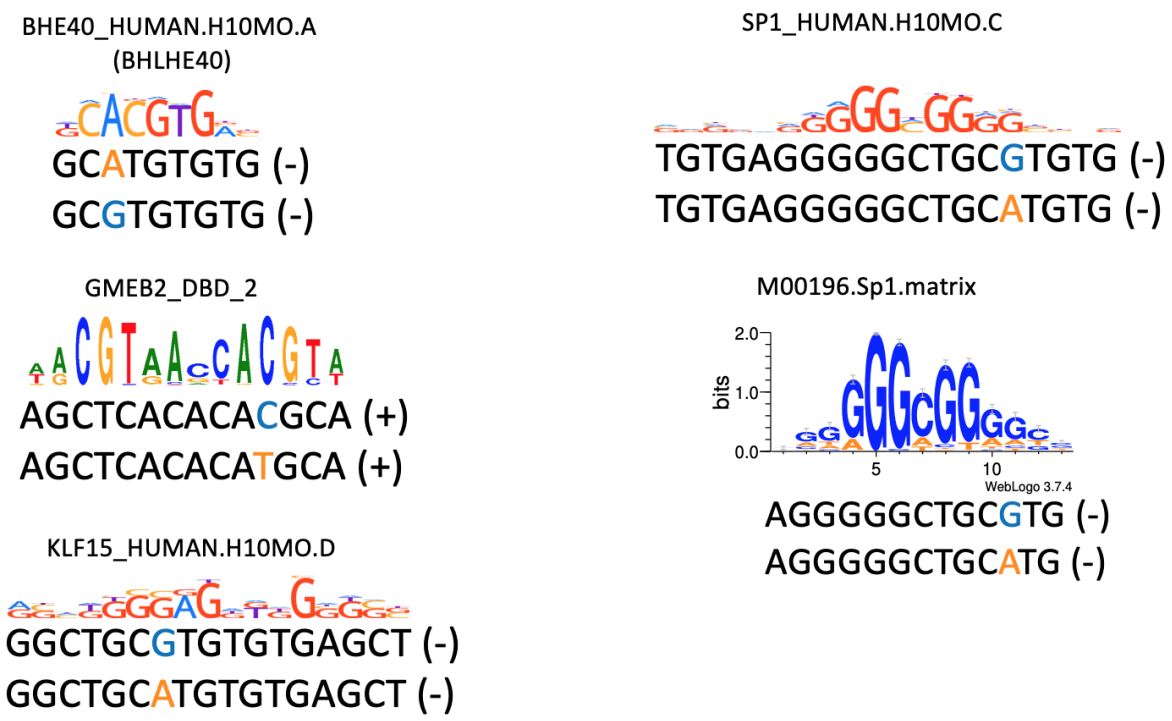


Figure C.6: (Figure C.4, C.5 continued) - motif logo of each candidate TF listed in Table C.6 and variants of the flanking sequences harboring rs67307131 risk allele T (A for - strand) or rs67307131 non-risk allele C (G for - strand). TFs are predicted to have higher binding affinity to the sequences on top than the sequences at bottom.

C.6 Other candidate TFs perturbed by rs12225399

motif from	motif to	motif id	motif source	TF	strand	hit allele (G/C)	Fimo *P*-value	permut *P*-value
118480277	118480289	M00491 .MAZR .matrix	TRANSFAC Human	PATZ1	-	C	3.98E-05	0.0092
118480281	118480299	IRX2_ HUMAN .H10MO.D	HOCOMOCO	IRX2	-	C	0.000942	0.00044
118480281	118480295	MA0732.1	JASPAR	EGR3	+	G	0.000732	0.0044
118480279	118480288	ZNF740 _full	jolma2013	ZNF740	+	C	0.000309	0.0025
118480278	118480288	ZN784_ HUMAN .H10MO.D	HOCOMOCO	ZNF784	-	C	0.000641	0.0065
118480277	118480290	AP2C_ HUMAN .H10MO.A	HOCOMOCO	TFAP2C	-	G	0.000777	0.0034
118480269	118480290	MAZ_ HUMAN .H10MO.A	HOCOMOCO	MAZ	-	C	0.00048	0.0094
118480266	118480287	MAZ_ HUMAN .H10MO.A	HOCOMOCO	MAZ	-	C	0.000171	0.013
118480284	118480290	M00624 .DBP .matrix	TRANSFAC Human	DBP	+	G	0.000763	0.017
118480271	118480287	MA0159.1	JASPAR	RXRA::RARA	-	C	0.000695	0.0067
118480276	118480287	ZN219_ HUMAN .H10MO.D	HOCOMOCO	ZNF219	-	C	0.00087	0.013
118480282	118480291	M01034 .Ebox .matrix	TRANSFAC Human	DEC1::TCF3::USF2:: TAL2::MASH1::MYC:: TCF4::MAD1::MYCN:: DEC2::TFEA::EHAND:: MYF5::HEN1::MYOG:: HAND2::USF1::MITF:: MXI1::MYF6::HTF4:: TFEB::MYOD::TCF12:: TAL1::MAX::MRF2	+	C	0.000821	0.016
118480280	118480289	THB_ HUMAN .H10MO.S	HOCOHOCO	THRB	+	G	0.000784	0.053
118480280	118480292	PKNX2_ HUMAN .H10MO.D	HOCOMOCO	PKNOX2	-	C	0.000114	0.055

Caption in the next page

Table C.8: Motifs of other candidate TFs perturbed by rs12225399. Columns from left to right: motif start coordinate in chr11 (hg19); motif end coordinate in chr11 (hg19); motif id from one of the four databases - JASPAR [72], HOCOMOCOv10 [73], TRANSFAC [74] and Jolma2013 [75]; motif source, one of the four databases - JASPAR [72], HOCOMOCOv10 [73], TRANSFAC [74] and Jolma2013 [75]; transcription factor name; strand harboring the motif; rs12225399 allele harbored by the motif; P-value from Fimo output; P-value from the permutation test.

Group	TF	TF avg expr	GG number	GC number	CC number	r_{GG}	r_{GC}	r_{CC}
All TCGA-LGG samples	PATZ1	10.85	281	166	29	0.14	0.24	0.45
IDH^{mut} only and TP	PATZ1	11.01	202	103	16	0.23	0.32	0.52
IDH^{mut} only	PATZ1	10.99	122	62	8	0.20	0.29	0.28
TP	PATZ1	11.03	80	41	8	0.24	0.37	0.70
All TCGA-LGG samples	IRX2	5.91	281	166	29	-0.07	0.09	0.32
IDH^{mut} only and TP	IRX2	6.49	202	103	16	-0.06	0.073	0.27
IDH^{mut} only	IRX2	7.33	122	62	8	0.01	0.20	-0.35
TP	IRX2	5.24	80	41	8	0.06	0.15	0.32
All TCGA-LGG samples	EGR3	8.34	281	166	29	-0.25	-0.03	0.11
IDH^{mut} only and TP	EGR3	8.26	202	103	16	-0.23	-0.11	0.065
IDH^{mut} only	EGR3	8.53	122	62	8	-0.11	-0.019	0.19
TP	EGR3	7.86	80	41	8	-0.31	-0.30	-0.68
All TCGA-LGG samples	ZNF740	9.82	281	166	29	0.16	0.25	0.093
IDH^{mut} only and TP	ZNF740	9.88	202	103	16	0.28	0.26	-0.13
IDH^{mut} only	ZNF740	9.85	122	62	8	0.22	0.17	-0.36
TP	ZNF740	9.94	80	41	8	0.29	0.45	-0.03
All TCGA-LGG samples	ZNF784	6.93	281	166	29	0.14	0.11	-0.18
IDH^{mut} only and TP	ZNF784	6.90	202	103	16	0.08	0.15	0.40
IDH^{mut} only	ZNF784	7.02	122	62	8	0.08	0.23	0.44
TP	ZNF784	6.73	80	41	8	0.19	0.19	0.16
All TCGA-LGG samples	TFAP2C	2.40	281	166	29	-0.11	-0.093	-0.014
IDH^{mut} only and TP	TFAP2C	2.21	202	103	16	-0.13	-0.23	-0.034
IDH^{mut} only	TFAP2C	2.39	122	62	8	-0.14	-0.14	-0.10
TP	TFAP2C	1.96	80	41	8	-0.03	-0.36	-0.41
All TCGA-LGG samples	MAZ	12.01	281	166	29	0.07	0.12	0.001
IDH^{mut} only and TP	MAZ	12.04	202	103	16	0.08	0.165	0.255
IDH^{mut} only	MAZ	11.98	122	62	8	-0.03	0.158	0.257
TP	MAZ	12.14	80	41	8	0.12	0.16	0.31
All TCGA-LGG samples	DBP	8.86	281	166	29	-0.25	0.015	-0.086
IDH^{mut} only and TP	DBP	8.88	202	103	16	-0.24	-0.11	0.35
IDH^{mut} only	DBP	9.02	122	62	8	-0.25	-0.11	0.37
TP	DBP	8.67	80	41	8	-0.11	-0.02	0.36
All TCGA-LGG samples	RXRA	11.04	281	166	29	-0.18	-0.04	-0.28
IDH^{mut} only and TP	RXRA	11.02	202	103	16	-0.10	-0.08	-0.32
IDH^{mut} only	RXRA	11.06	122	62	8	-0.11	0.017	-0.31
TP	RXRA	10.96	80	41	8	-0.05	-0.17	-0.31
All TCGA-LGG samples	RARA	9.73	281	166	29	0.23	0.14	0.15
IDH^{mut} only and TP	RARA	9.78	202	103	16	0.24	0.28	0.40
IDH^{mut} only	RARA	9.68	122	62	8	0.11	0.28	0.41
TP	RARA	9.91	80	41	8	0.33	0.25	0.59
All TCGA-LGG samples	ZNF219	9.97	281	166	29	0.30	0.36	0.18
IDH^{mut} only and TP	ZNF219	10.05	202	103	16	0.29	0.38	0.11
IDH^{mut} only	ZNF219	9.82	122	62	8	0.12	0.27	-0.21
TP	ZNF219	10.39	80	41	8	0.36	0.45	0.78

Continue in the next page

Group	TF	TF avg expr	GG number	GC number	CC number	r_{GG}	r_{GC}	r_{CC}
IDH^{mut} only and TP	MYC	10.27	202	103	16	0.22	0.38	0.44
IDH^{mut} only	MYC	10.29	122	62	8	0.16	0.31	0.46
TP	MYC	10.25	80	41	8	0.35	0.52	0.17
All TCGA-LGG samples	TCF12	12.89	281	166	29	0.15	0.27	0.45
IDH^{mut} only and TP	TCF12	13.15	202	103	16	0.20	0.29	0.22
IDH^{mut} only	TCF12	13.06	122	62	8	0.25	0.26	0.36
TP	TCF12	13.27	80	41	8	0.06	0.29	0.19
All TCGA-LGG samples	TCF3	10.82	281	166	29	0.16	0.18	0.054
IDH^{mut} only and TP	TCF3	10.89	202	103	16	0.17	0.30	0.30
IDH^{mut} only	TCF3	10.96	122	62	8	0.19	0.30	0.12
TP	TCF3	10.79	80	41	8	0.23	0.39	0.25
All TCGA-LGG samples	TFEB	9.62	281	166	29	0.52	0.55	0.58
IDH^{mut} only and TP	TFEB	9.66	202	103	16	0.50	0.42	0.67
IDH^{mut} only	TFEB	9.70	122	62	8	0.53	0.40	0.76
TP	TFEB	9.60	80	41	8	0.50	0.54	0.48
All TCGA-LGG samples	USF1	10.23	281	166	29	0.14	0.10	-0.19
IDH^{mut} only and TP	USF1	10.25	202	103	16	0.12	0.12	0.19
IDH^{mut} only	USF1	10.24	122	62	8	0.11	0.14	0.57
TP	USF1	10.27	80	41	8	0.10	0.12	-0.12
All TCGA-LGG samples	USF2	11.53	281	166	29	-0.025	0.05	-0.05
IDH^{mut} only and TP	USF2	11.51	202	103	16	-0.08	-0.016	0.36
IDH^{mut} only	USF2	11.80	122	62	8	0.06	0.09	0.51
TP	USF2	11.06	80	41	8	0.16	0.15	0.12
All TCGA-LGG samples	ZEB2	11.82	281	166	29	0.39	0.42	0.54
IDH^{mut} only and TP	ZEB2	11.89	202	103	16	0.45	0.38	0.15
IDH^{mut} only	ZEB2	11.82	122	62	8	0.53	0.23	0.06
TP	ZEB2	11.99	80	41	8	0.29	0.57	0.14
All TCGA-LGG samples	THRB	8.91	281	166	29	-0.29	-0.12	-0.24
IDH^{mut} only and TP	THRB	8.87	202	103	16	-0.25	-0.21	-0.72
IDH^{mut} only	THRB	8.93	122	62	8	-0.26	-0.20	-0.66
TP	THRB	9.07	80	41	8	-0.32	-0.41	-0.77
All TCGA-LGG samples	PKNOX2	9.41	281	166	29	0.04	0.23	0.31
IDH^{mut} only and TP	PKNOX2	9.53	202	103	16	0.04	0.22	0.11
IDH^{mut} only	PKNOX2	9.49	122	62	8	0.00	0.26	0.01
TP	PKNOX2	9.60	80	41	8	0.05	0.081	0.001

Table C.9: The Pearson's correlation coefficient (r) between TF and *PHLDB1* stratified into rs12225399 GG, GC and CC genotype groups. Columns from left to right: patient group; transcription factor name; TF average expression, defined as $\overline{log_2(\text{RSEM}+1)}$; patient number in GG genotype group; patient number in GC genotype group; patient number in CC genotype group; Pearson's correlation coefficient in GG genotype group; Pearson's correlation coefficient in GC genotype group; Pearson's correlation coefficient in CC genotype group.

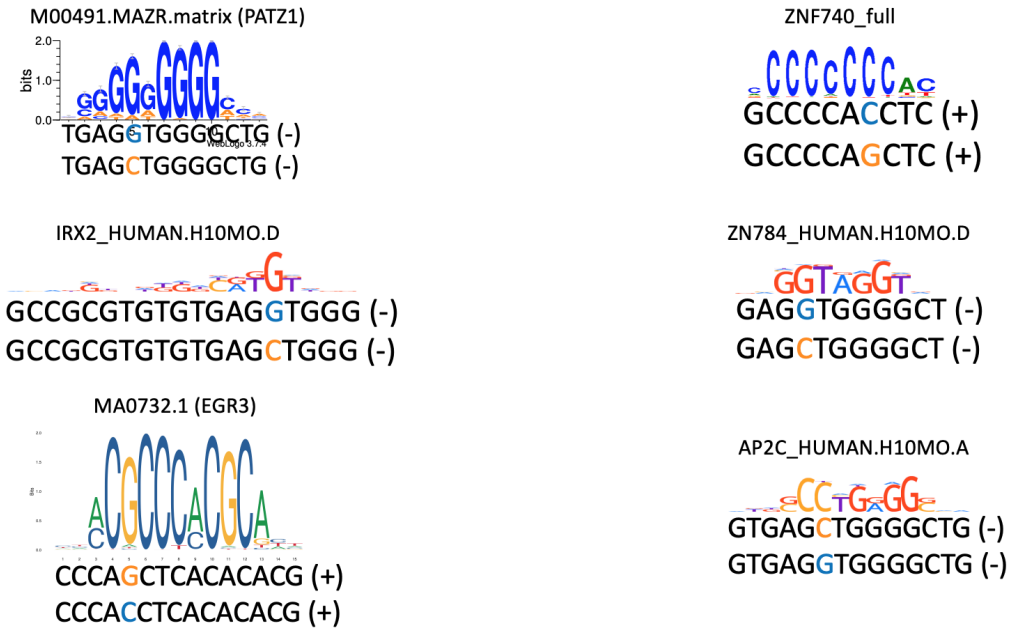


Figure C.7: Motif logo of each candidate TF listed in Table C.8 and variants of the flanking sequences harboring rs12225399 risk allele G (C for - strand) or rs12225399 non-risk allele C (G for - strand). TFs are predicted to have higher binding affinity to the sequences on top than the sequences at bottom.

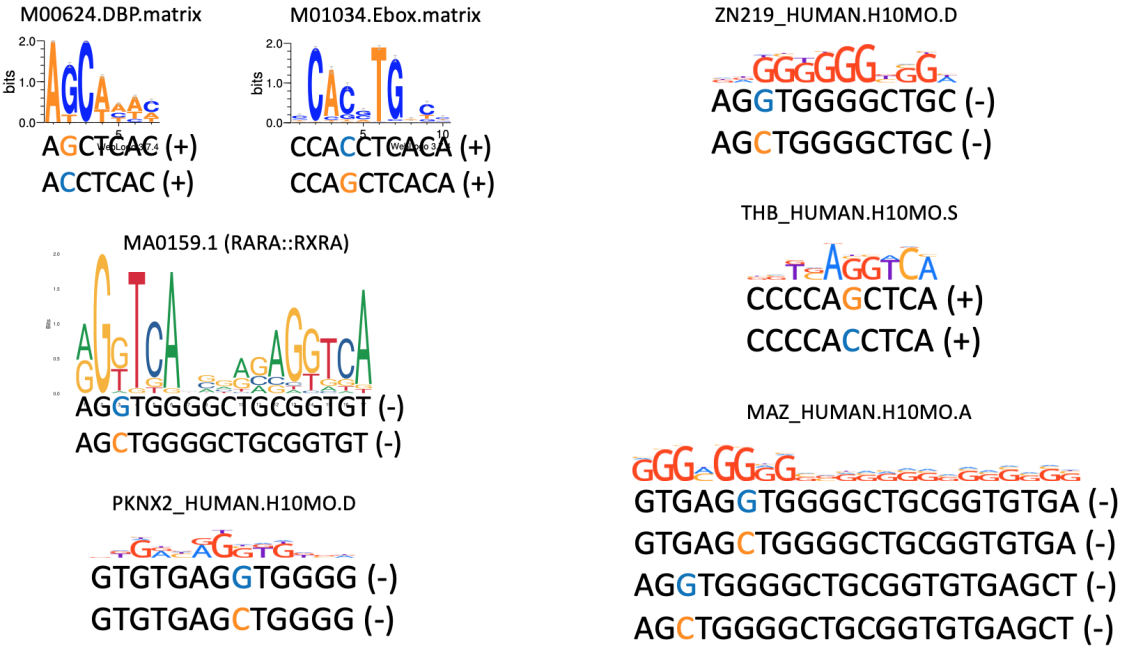


Figure C.8: (Figure C.7 continued) - motif logo of each candidate TF listed in Table C.8 and variants of the flanking sequences harboring rs12225399 risk allele G (C for - strand) or rs12225399 non-risk allele C (G for - strand). TFs are predicted to have higher binding affinity to the sequences on top than the sequences at bottom.

C.7 Summary of DNase-seq and ChIP-seq files in the CNN model

Cell line	DNase-seq peak files	DNase-seq bigwig files	SP1 ChIP-seq optimal peak files	SP1 ChIP-seq raw peak files	SP1 ChIP-seq bigwig files
H1-hESC	ENCFF807SXP	ENCFF899UPI	ENCFF317TBZ	ENCFF721YVP	ENCFF433IIE
HEK293T	ENCFF910QHN	ENCFF716SFD	ENCFF240PYU	ENCFF730EIH	ENCFF306LPM
HepG2	ENCFF571RHF	ENCFF867UYB	ENCFF573ALP	ENCFF894RIX	ENCFF732DBE
Liver	ENCFF286LYP	ENCFF219YEB	ENCFF017YUI	ENCFF027RDL	ENCFF672UVA
K562	ENCFF623AFX	ENCFF789QUY	ENCFF300XUA	ENCFF103ESU	ENCFF591WIF
MCF-7	ENCFF382GKE	ENCFF216EVD	ENCFF193MKL	ENCFF546ASI	ENCFF475AXY
A549	ENCFF475BVB	ENCFF723TWJ	ENCFF348RKC	ENCFF831GEW	ENCFF715NIC

Table C.10: DNase-seq and ChIP-seq files from ENCODE used in CNN model training

DNase-seq wig files
GSM1027328_UW.Fetal_Brain.ChromatinAccessibility.H-24510.DNase.DS20780.wig
GSM530651_UW.Fetal_Brain.ChromatinAccessibility.H-22510.DS11872.wig
GSM595913_UW.Fetal_Brain.ChromatinAccessibility.H-22510.DS11877.wig
GSM595920_UW.Fetal_Brain.ChromatinAccessibility.H-22911.DS14464.wig
GSM595922_UW.Fetal_Brain.ChromatinAccessibility.H-23266.DS14717.wig
GSM595923_UW.Fetal_Brain.ChromatinAccessibility.H-23266.DS14718.wig
GSM595926_UW.Fetal_Brain.ChromatinAccessibility.H-23284.DS14803.wig
GSM595928_UW.Fetal_Brain.ChromatinAccessibility.H-23284.DS14815.wig
GSM665804_UW.Fetal_Brain.ChromatinAccessibility.H-23399.DS15453.wig
GSM665819_UW.Fetal_Brain.ChromatinAccessibility.H-23548.DS16302.wig
GSM878650_UW.Fetal_Brain.ChromatinAccessibility.H-24279.DS20221.wig
GSM878651_UW.Fetal_Brain.ChromatinAccessibility.H-24297.DS20226.wig
GSM878652_UW.Fetal_Brain.ChromatinAccessibility.H-24381.DS20231.wig

Table C.11: Fetal Brain DNase-seq raw signal files from REMC

Appendix D

Supplementary Material for Chapter 5

D.1 GWAS SNP rs11706832 and its high LD SNPs

RS number	Chromosome	Position (hg19)	Alleles	r^2	GWAS SNP correlated alleles
rs11706832	3	66502981	A/C	1	-
rs56300148	3	66497714	T/C	0.93	A=T, C=C
rs4402869	3	66507444	G/A	0.87	A=G, C=A
rs11717516	3	66508132	G/A	0.86	A=G, C=A

Table D.1: GWAS SNP rs11706832 and its three high LD ($r^2 \geqslant 0.8$, 1000 Genomes Phase 3, EUR) SNPs.

D.2 LEF1 motif and its expression correlation with *SLC25A26*

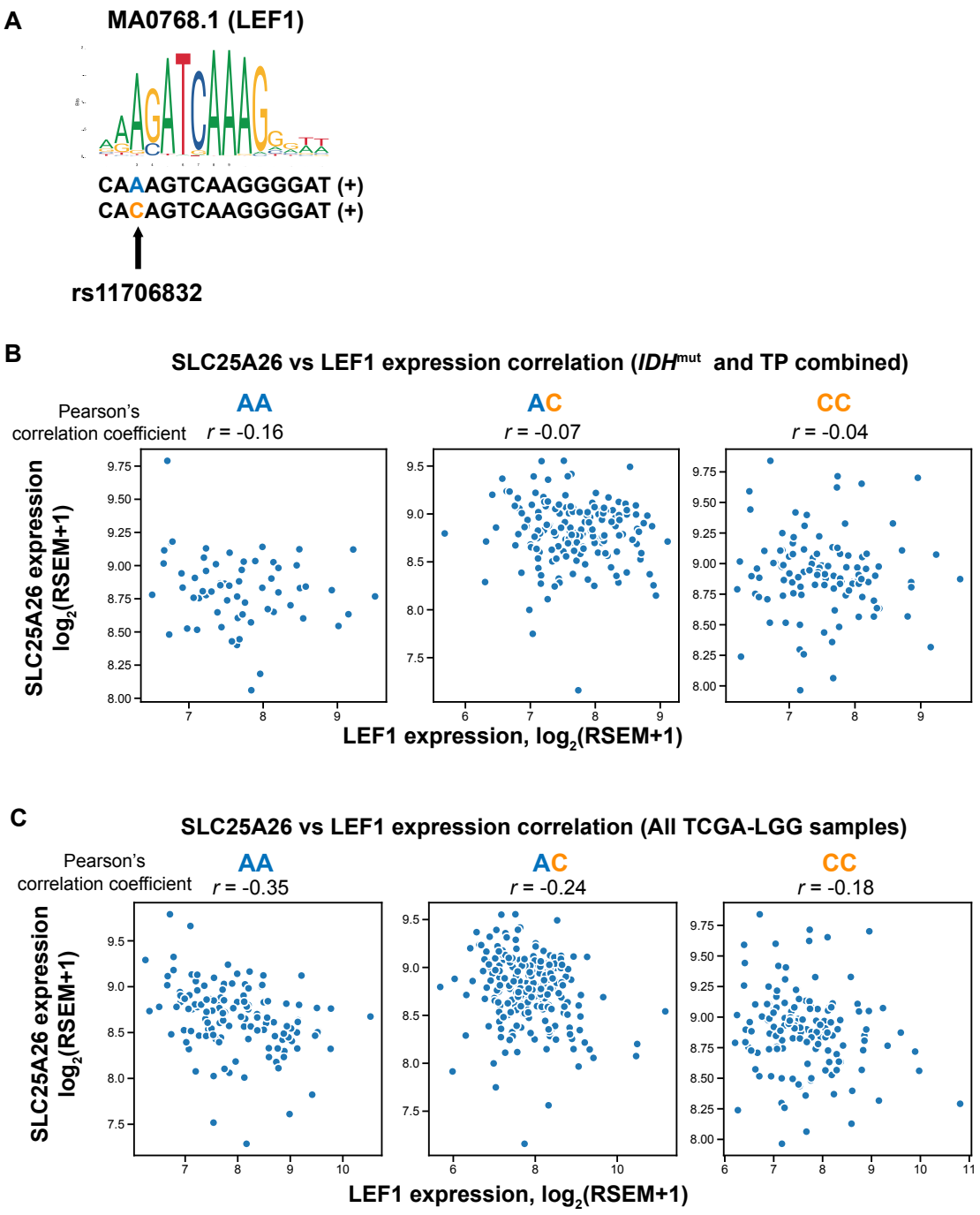


Figure D.1: The GWAS SNP rs11706832 likely modulates *SLC25A26* expression through perturbing the binding affinity of LEF1. (A) LEF1 motif MA0768.1 (JASPAR [24]) and two versions of the flanking sequence harboring the rs11706832-A and rs11706832-C alleles. (B) Scatter plots of *LEF1* vs. *SLC25A26* expression in the three genotypes of rs11706832 in the combined TCGA-LGG "IDH^{mut} only" and triple-positive group. (C) Scatter plots of *LEF1* vs. *SLC25A26* expression in the three genotypes of rs11706832 in all TCGA-LGG samples.

D.3 Other candidate TFs perturbed by rs11706832

motif from	motif to	motif id	motif source	TF	strand	hit allele (A/C)	Fimo P-value	permut P-value
66502978	66502990	COT2_HUMAN.H10MO.A	HOCOMOCO	NR2F2	+	A	0.000646	0.0088
66502980	66502994	COT2_HUMAN.H10MO.S	HOCOMOCO	NR2F2	+	A	0.000568	0.0099
66502978	66502988	MA0141.3	JASPAR	ESRRB	+	A	0.000427	0.0025
66502970	66502983	NFATC1_full_3	jolma2013	NFATC1	-	A	0.000545	0.0062
66502970	66502983	NFATC1_full_3	jolma2013	NFATC1	+	A	0.000522	0.013
66502980	66502992	NR6A1_HUMAN.H10MO.B	HOCOMOCO	NR6A1	+	A	0.00082	0.037

Table D.2: Motifs of other candidate TFs perturbed by rs11706832. Columns from left to right: motif start coordinate in chr3 (hg19); motif end coordinate in chr3 (hg19); motif id from one of the four databases - JASPAR [72], HOCOMOCOv10 [73], TRANSFAC [74] and Jolma2013 [75]; motif source, one of the four databases - JASPAR [72], HOCOMOCOv10 [73], TRANSFAC [74] and Jolma2013 [75]; transcription factor name; strand harboring the motif; rs11706832 allele harbored by the motif; P-value from Fimo output; P-value from the permutation test.

Group	TF	TF avg expr	AA number	AC number	CC number	r_{AA}	r_{AC}	r_{CC}
All TCGA-LGG samples	NR2F2	7.98	132	232	143	-0.31	-0.29	-0.12
IDH^{mut} only and TP	NR2F2	7.73	64	167	109	-0.16	-0.16	-0.073
IDH^{mut} only	NR2F2	7.88	39	95	69	0.007	-0.11	-0.07
TP	NR2F2	7.51	25	72	40	-0.42	-0.25	-0.06
All TCGA-LGG samples	ESRRB	2.98	132	232	143	-0.26	-0.20	-0.10
IDH^{mut} only and TP	ESRRB	2.86	64	167	109	-0.30	-0.16	-0.06
IDH^{mut} only	ESRRB	3.02	39	95	69	-0.41	-0.25	-0.12
TP	ESRRB	2.63	25	72	40	0.02	0.048	0.05
All TCGA-LGG samples	NFATC1	7.49	132	232	143	-0.42	-0.15	-0.10
IDH^{mut} only and TP	NFATC1	7.31	64	167	109	-0.25	-0.006	-0.11
IDH^{mut} only	NFATC1	7.99	39	95	69	-0.26	-0.034	-0.17
TP	NFATC1	6.30	25	72	40	-0.09	0.07	-0.045
All TCGA-LGG samples	NR6A1	2.81	132	232	143	-0.05	-0.09	-0.3
IDH^{mut} only and TP	NR6A1	2.81	64	167	109	-0.31	-0.08	-0.26
IDH^{mut} only	NR6A1	3.00	39	95	69	-0.28	-0.14	-0.17
TP	NR6A1	2.53	25	72	40	-0.27	0.01	-0.33

Caption on the next page.

Table D.3: The Pearson's correlation coefficient (r) between TF and *SLC25A26* stratified into rs11706832 AA, AC and CC genotype groups. Columns from left to right: patient group; transcription factor name; TF average expression, defined as $\overline{log_2(\text{RSEM}+1)}$; patient number in AA genotype group; patient number in AC genotype group; patient number in CC genotype group; Pearson's correlation coefficient in AA genotype group; Pearson's correlation coefficient in AC genotype group; Pearson's correlation coefficient in CC genotype group.

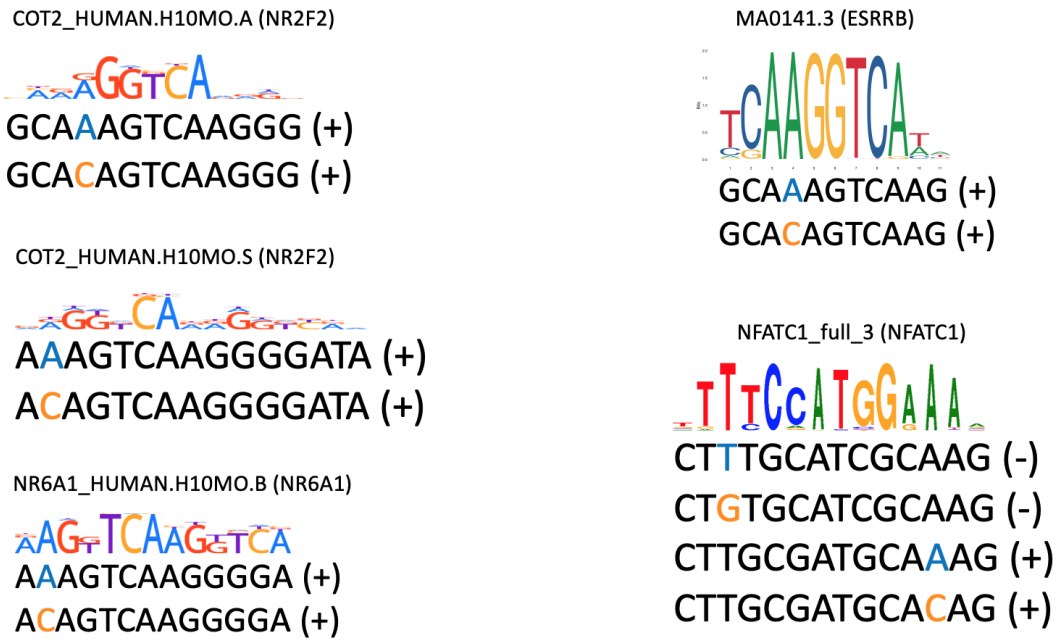


Figure D.2: Motif logo of each candidate TF listed in Table D.2 and variants of the flanking sequences harboring rs11706832 risk allele C (G for - strand) or rs11706832 non-risk allele A (T for - strand). TFs are predicted to have higher binding affinity to the sequences on top than the sequences at bottom.

D.4 Candidate TFs perturbed by rs4402869

motif from	motif to	motif id	motif source	TF	strand	hit allele (G/A)	Fimo *P*-value	permut *P*-value
66507429	66507447	P63_HUMAN .H10MO.A	HOCO-MOCO	TP63	-	G	5.04E-05	0.0049
66507429	66507446	MA0525.2	JASPAR	TP63	-	G	0.000383	0.013
66507443	66507454	M01037 .GLI.matrix	TRANSF-AC Human	GLI3 ::GLI2 ::GLI	-	G	0.000803	0.024
66507432	66507451	MA0066.1	JASPAR	PPARG	+	G	0.000366	0.033
66507433	66507449	ESR1_DBD	jolma2013	ESR1	-	G	0.000944	0.017
66507443	66507455	MA0815.1	JASPAR	TFAP2C	+	G	0.00099	0.01
66507443	66507455	MA0815.1	JASPAR	TFAP2C	-	G	0.00099	0.01

Table D.4: Motifs of candidate TFs perturbed by rs4402869. Columns from left to right: motif start coordinate in chr3 (hg19); motif end coordinate in chr3 (hg19); motif id from one of the four databases - JASPAR [72], HOCOMOCOv10 [73], TRANSFAC [74] and Jolma2013 [75]; motif source, one of the four databases - JASPAR [72], HOCOMOCOv10 [73], TRANSFAC [74] and Jolma2013 [75]; transcription factor name; strand harboring the motif; rs4402869 allele harbored by the motif; *P*-value from Fimo output; *P*-value from the permutation test.

Group	TF	TF avg expr	GG number	GA number	AA number	r_{GG}	r_{GA}	r_{AA}
All TCGA-LGG samples	TP63	2.50	133	212	120	-0.225	-0.26	-0.28
IDH^{mut} only and TP	TP63	2.24	68	156	91	-0.25	-0.28	-0.18
IDH^{mut} only	TP63	2.20	41	93	57	-0.21	-0.25	-0.19
TP	TP63	2.31	27	63	34	-0.38	-0.33	-0.18
All TCGA-LGG samples	GLI2	7.09	133	212	120	-0.23	-0.091	-0.24
IDH^{mut} only and TP	GLI2	6.94	68	156	91	0.18	0.046	-0.22
IDH^{mut} only	GLI2	7.23	41	93	57	0.45	-0.058	-0.007
TP	GLI2	6.50	27	63	34	-0.23	0.22	-0.42
All TCGA-LGG samples	GLI3	7.87	133	212	120	-0.51	-0.077	-0.30
IDH^{mut} only and TP	GLI3	7.60	68	156	91	-0.24	-0.009	-0.32
IDH^{mut} only	GLI3	8.04	41	93	57	0.21	-0.05	-0.23
TP	GLI3	6.93	27	63	34	-0.50	0.018	-0.42
All TCGA-LGG samples	PPARG	5.01	133	212	120	-0.30	-0.13	0.075
IDH^{mut} only and TP	PPARG	4.83	68	156	91	-0.11	-0.063	0.23
IDH^{mut} only	PPARG	4.98	41	93	57	-0.077	-0.054	0.27
TP	PPARG	4.61	27	63	34	-0.14	-0.09	0.23
All TCGA-LGG samples	ESR1	3.5	133	212	120	-0.20	-0.16	-0.078
IDH^{mut} only and TP	ESR1	3.39	68	156	91	-0.18	-0.079	-0.025
IDH^{mut} only	ESR1	3.33	41	93	57	-0.18	0.019	-0.032
TP	ESR1	3.48	27	63	34	-0.22	-0.23	-0.017
All TCGA-LGG samples	TFAP2C	2.39	133	212	120	-0.29	0.024	0.013
IDH^{mut} only and TP	TFAP2C	2.23	68	156	91	-0.15	0.015	0.078
IDH^{mut} only	TFAP2C	2.42	41	93	57	0.03	0.035	0.27
TP	TFAP2C	1.94	27	63	34	-0.38	-0.034	-0.11

Table D.5: The Pearson's correlation coefficient (r) between TF and *SLC25A26* stratified into rs4402869 GG, GA and AA genotype groups. Columns from left to right: patient group; transcription factor name; TF average expression, defined as $\overline{log_2(\text{RSEM} + 1)}$; patient number in GG genotype group; patient number in GA genotype group; patient number in AA genotype group; Pearson's correlation coefficient in GG genotype group; Pearson's correlation coefficient in GA genotype group; Pearson's correlation coefficient in AA genotype group.

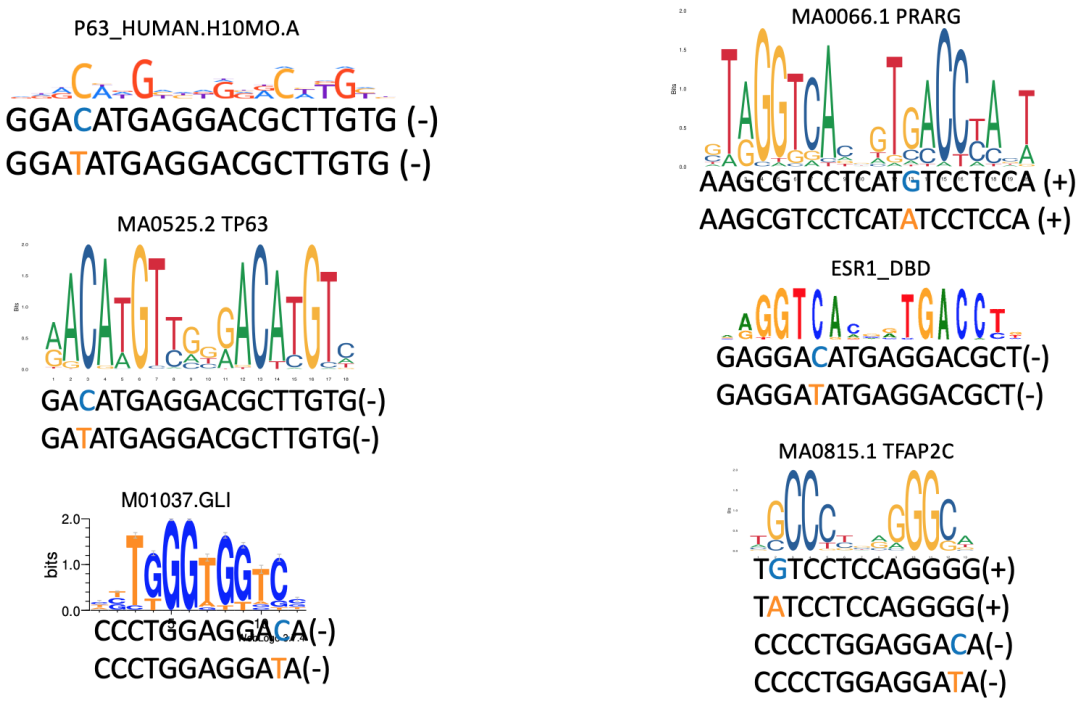


Figure D.3: Motif logo of each candidate TF listed in Table D.4 and variants of the flanking sequences harboring rs4402869 risk allele A (T for - strand) or rs4402869 non-risk allele G (C for - strand). TFs are predicted to have higher binding affinity to the sequences on top than the sequences at bottom.

Appendix E

Supplementary Material for Chapter 6

E.1 Related theorems and corollaries

The theorems and corollaries listed below (Theorem E.1.1, Theorem E.1.2, Corollary E.1.2.1, Corollary E.1.2.2) are from I. V. Oseledets, "Tensor-train decomposition" [34].

Theorem E.1.1. *If for each unfolding matrix A_k of a d-dimensional tensor* $\mathbf{A}$

$$rank(A_k) = r_k, \tag{E.1}$$

then there exists a decomposition (equation 6.2) with TT-ranks not higher than r_k.

Theorem E.1.2. *Suppose that the unfoldings A_k of the tensor* $\mathbf{A}$ *satisfy: $A_k = R_k + Q_k$, $rank(R_k) = r_k$, $\|Q_k\|_F = \varepsilon_k$, $k = 1, \ldots, d-1$. Then TT-SVD computes a tensor* $\mathbf{B}$ *in the TT-format with TT-ranks r_k and*

$$\|\mathbf{A} - \mathbf{B}\|_F \leq \sqrt{\sum_{k=1}^{d-1} \varepsilon_k^2}. \tag{E.2}$$

Corollary E.1.2.1. *If a tensor* $\mathbf{A}$ *admits a canonical approximation with R terms and accuracy ε, then there exists a TT-approximation with TT-ranks $r_k \leq R$ and accuracy $\sqrt{d-1}\varepsilon$.*

Corollary E.1.2.2. *Given a tensor* $\mathbf{A}$ *and rank bounds r_k, the best approximation to* $\mathbf{A}$ *in the Frobenius norm with TT-ranks bounded by r_k always exists (denote it by $\mathbf{A}^{best}$), and the TT-approximation* $\mathbf{B}$ *computed by the TT-SVD algorithm is quasi-optimal:*

$$\|\mathbf{A} - \mathbf{B}\|_F \leq \sqrt{d-1}\|\mathbf{A} - \mathbf{A}^{best}\|_F. \tag{E.3}$$

E.2 Supplementary figures

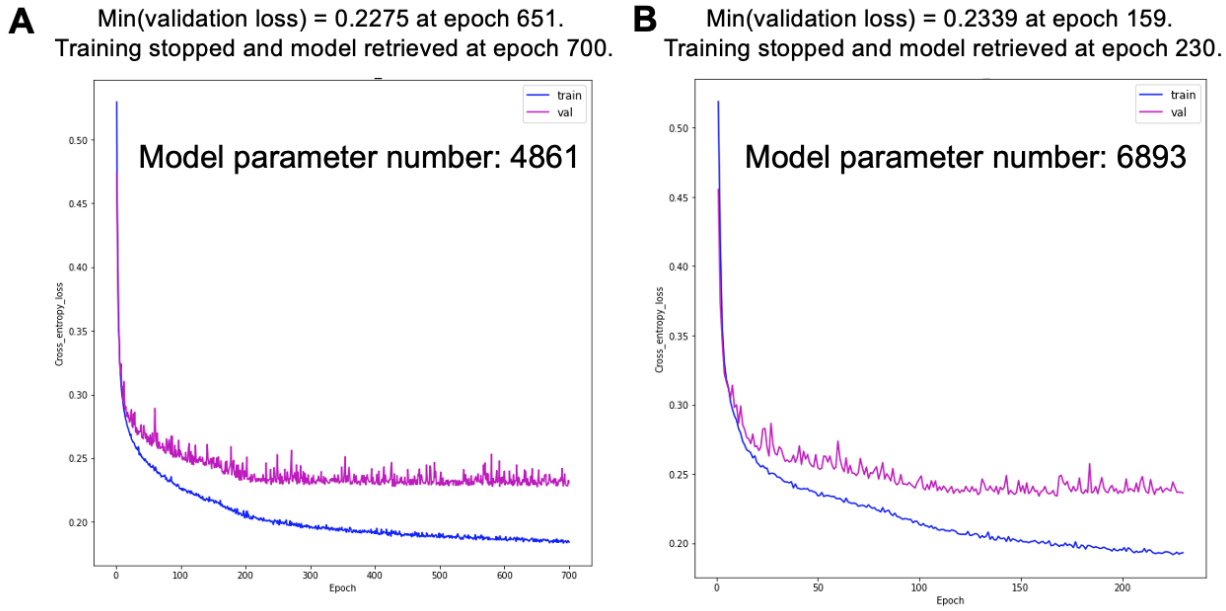


Figure E.1: The cross entropy loss vs epochs of each SP1 CNN-TT model configuration. (A) The cross entropy loss vs epochs of the CNN-TT model with configuration $(r_0, r_1, r_2, r_3) = (1, 4, 4, 1)$ and $(m_1, m_2, m_3) \times (n_1, n_2, n_3) = (4, 4, 5) \times (4, 4, 60)$. The total parameter number is 4861. The training was stopped and the trained model was retrieved at epoch 700. (B) Similar to (A), but for configuration $(r_0, r_1, r_2, r_3) = (1, 8, 8, 1)$ and $(m_1, m_2, m_3) \times (n_1, n_2, n_3) = (4, 4, 5) \times (4, 4, 60)$. The total parameter number is 6893. The training was stopped and the trained model was retrieved at epoch 230.

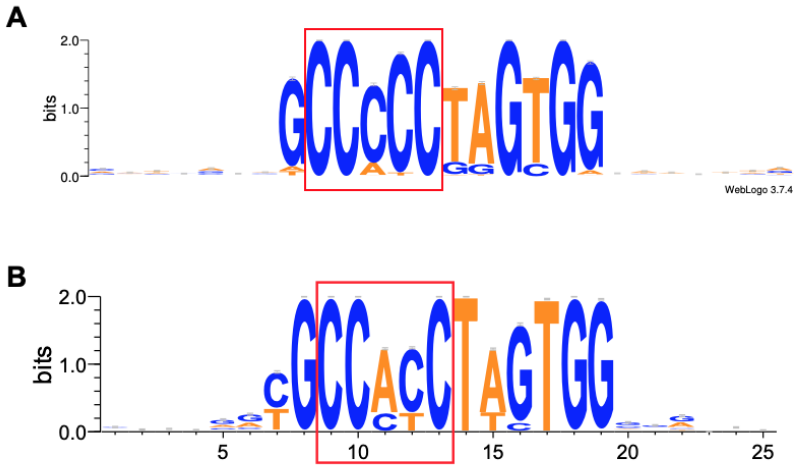


Figure E.2: The core motif learned by the CNN-TT model resembled the core motif of SP1 MA0079.3. (A) One of the motifs learned by the CNN-TT model with configuration $(r_0, r_1, r_2, r_3) = (1, 8, 8, 1)$ and $(m_1, m_2, m_3) \times (n_1, n_2, n_3) = (4, 4, 5) \times (4, 4, 60)$ (total parameter number 6893), visualized through a motif logo obtained from WebLogo [79] 3. The core motif inside the red box resembles the core motif of SP1 MA0079.3 (Figure 4.3A). (B) The motif learned by the original CNN model in Section 4.3. The learned motif from the CNN-TT model resembles the learned motif from the original CNN model.

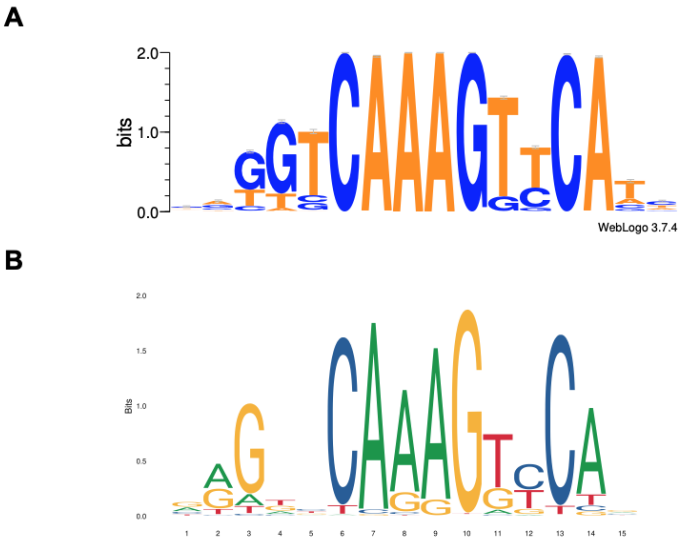


Figure E.3: Another motif learned by the CNN-TT model resembles the HNF4A motif MA0114.2. (A) Another motif learned by the CNN-TT model with configuration $(r_0, r_1, r_2, r_3) = (1, 8, 8, 1)$ and $(m_1, m_2, m_3) \times (n_1, n_2, n_3) = (4, 4, 5) \times (4, 4, 60)$ (total parameter number 6893), visualized through a motif logo obtained from WebLogo [79] 3. (B) The motif logo of HNF4A MA0114.2 from the JASPAR [24] database.